Modeling Non-Stationarities in High-Frequency Financial Time Series

Linda Ponta[1Y], Mailan Trinh[2Y], Marco Raberto[1Y], Enrico Scalas[2,3Y,*], Silvano Cincotti[1Y],

1 DIME-CINEF, University of Genoa, Genoa, Italy
2 Department of Mathematics, School of Mathematical and Physical Sciences, University of Sussex, Brighton, UK
3 BCAM - Basque Center for Applied Mathematics, Bilbao, Basque Country - Spain

Y These authors contributed equally to this work.
*e.scalas@sussex.ac.uk

Abstract

We study tick-by-tick financial returns belonging to the FTSE MIB index of the Italian Stock Exchange (Borsa Italiana). We can confirm previously detected non-stationarities. However, scaling properties reported in the previous literature for other high-frequency financial data are only approximately valid. As a consequence of the empirical analyses, we propose a simple method for describing non-stationary returns, based on a non-homogeneous normal compound Poisson process. We test this model against the empirical findings and it turns out that the model can approximately reproduce several stylized facts of high-frequency financial time series. Moreover, using Monte Carlo simulations, we analyze order selection for this model class using three information criteria: Akaike′s information criterion (AIC), the Bayesian information criterion (BIC) and the Hannan-Quinn information criterion (HQ). For comparison, we also perform a similar Monte Carlo experiment for the ACD (autoregressive conditional duration) model. Our results show that the information criteria work best for small parameter numbers for the compound Poisson type models, whereas for the ACD model the model selection procedure does not work well in certain cases.

Introduction

The recent rise in the availability of high-frequency financial data has seen an increase in the number of studies focusing on the areas of classification and modeling of financial markets at the ultra-high frequency level. The development of models that are able to reflect the various phenomena observed in real data is an important step towards obtaining a full understanding of the fundamental stochastic processes driving the market. The statistical properties of high-frequency financial data and market micro-structural properties were studied by means of different tools, including phenomenological models of price dynamics and agent-based market simulations (see [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14] [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31]).

Various studies on high-frequency econometrics appeared in the literature using the autoregressive conditional duration (ACD) models (see [32], [33], [34], [35]).

Alternative stochastic models were also proposed, e.g., diffusive models, ARCH-GARCH models, stochastic volatility models, models based on fractional processes, models based on subordinate processes (see [36], [37], [38], [39], [40], [41], [42]) as well as models based on self-exciting processes of Hawkes type [43], [44], [45]. An important variable is the order imbalance. Many existing studies analyze order imbalances around specific events or over short periods of time. For example, in [46] order imbalances are analyzed around the October 1987 crash. [47] analyzes how order imbalances change the relation between stock volatility and volume using data for about six months. A large body of research examines the effect of the bid-ask spread and the order impact on the short-run behavior of prices (see [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61]). Trading activity was measured by the average number of trades in unit time intervals in [62] and [63]. However, aggregating trades into time intervals of the same length may have influences on the analysis. For instance, if intervals are too short with respect to the average waiting time between consecutive trades then every interval will contain either no point or few points. On the contrary, if intervals are too long, aggregation of too many points may lead to loss of information on the time structure of the process. Moreover, in both cases one distorts the kurtosis of the return process (see [33]).

For this reason, an important empirical variable is the waiting time between two consecutive transactions (see [10], [64], [65], [25], [21], [22], [23], [24]). In the market, during a trading day the activity is not constant (see [32], [33]) leading to fractal-time behavior (see [66], [67]). Indeed, as a consequence of the double auction mechanism, waiting times between two subsequent trades are themselves random variables (see [64], [68], [69]). They may also be correlated to returns (see [70]) as well as to traded volumes. In the last few years in order to investigate tick-by-tick financial time series, the continuous-time random walk (CTRW) was used (see [4], [71], [72], [64]). It turned out that interorder and intertrade waiting-times are not exponentially distributed. Therefore, the jump process of tick-by-tick prices is non-Markovian (see [4], [64]). Bianco and Grigolini applied a new method to verify whether the intertrade waiting time process is a genuine renewal process (see [73], [74], [75]). This was assumed by the CTRW hypothesis in [4]. They found that intertrade waiting-times do follow a renewal process. Indeed, trading via the order book is asynchronous and a transaction occurs only if a trader issues a market order. For liquid stocks, waiting times can vary in a range between fractions of a second to a few minutes, depending on the specific stock and on the market considered. In [70], the reader can find a study on General Electric stocks traded in October 1999. Waiting times between consecutive prices exhibit 1-day periodicity, typical of variable intraday market activity. Moreover, the survival probability (the complementary cumulative distribution function) of waiting times is not exponentially distributed (see [76], [64]), but is well fitted by a Weibull function (see [77], [32], [33], [70], [14]).

Here, inspired by [78], we propose a model based on non-homogeneous Poisson processes. The paper is organized as follows. Section 1 describes the data set. A general description is presented in Subsection 1.1, and the FTSE MIB index in Subsection 1.2. Section 2 (and in particular Subsections 2.1 and 2.2) describes the statistical analysis of the single assets and of the FTSE MIB index, respectively as well as the scaling analysis; Section 3 contains the bivariate analysis whereas Section 4 is devoted to the compound Poisson model, its order selection and the numerical results. A comparison with order selection performance for ACD models is presented in the same section. Finally, Section 5 presents the conclusions of this work.

1 Data set

1.1 General description

The data set includes high-frequency trades registered at Italian Stock Exchange (BIt or Borsa Italiana), from the 03rd of February 2011 to the 09th of March 2011. The data of February 14th 2011 are not used because, on that day, there were technical problems at BIt. Moreover, we have removed the data of the 21st of February, as well. In fact, on that day, there was a crash in the Italian market related to the events in Lybia (on the 15th of February, a rebellion against the Lybian government begun). We consider the 40 shares in the FTSE MIB index at the time, namely: A2A, STS, ATL, AGL, AZM, BP, BMP, PMI, BUL, BZU, CPR, DIA, ENE, EGP, ENI, EXO, F, FI, FNC, FSA, G, IPG, ISP, LTO, LUX, MS, MB, MED, PLT, PC, PRY, SPM, SRG, STM, TIT, TEN, TRN, TOD, UBI, UCG. Further information on the database and the full meaning of the symbols is available from www.borsaitaliana.it . Table 1 shows the meaning of the ticker symbols as well as the number of observations for each share. The forty stocks composing the FTSE MIB vary in their average market capitalization and exhibit different levels of trading activity with different numbers of trades over this period as summarized in the last column in Table 1 where the total number of observations in the chosen month is given. Choosing one month of high-frequency data was a trade-off between the necessity of using enough data for significant statistical analysis and, on the other hand, the goal of minimizing the effect of external economic fluctuations leading to non-stationarities of the kind discussed in [79]. For every stock, the data set consists of prices $p(t_i)$, volumes $v(t_i)$ and times of execution t_i, where i is the trade index, varying from 1 to the total number of daily trades N. These data were filtered in order to remove misprints in prices and times of execution. In particular, concerning prices, when there are multiple prices for the same time of execution, we consider only one transaction at that time and a price equal to the average of the multiple prices, and concerning the waiting times, τ, between two executions, we remove observations larger than 200 s: This means more than 3 minutes without recorded trading.

1.2 FTSE MIB Index

The FTSE MIB Index (see [80]) is the primary benchmark index for the Italian equity markets. Capturing approximately 80% of the domestic market capitalisation, the Index is made up of highly liquid, leading companies across Industry Classification Benchmark (ICB) sectors in Italy. The FTSE MIB Index measures the performance of 40 shares listed on Borsa Italiana and seeks to replicate the broad sector weights of the Italian stock market. The Index is derived from the universe of stocks trading on BIt. The Index replaces the previous S&P/MIB Index, as a benchmark Index for Exchange Traded Funds (ETFs), and for tracking large capitalisation stocks in the Italian market. FTSE MIB Index is calculated on a real-time basis in EUR. The official opening and closing hours of the FTSE MIB Index series coincide with those of BIt markets and are 09:01 and 17:31 respectively. The FTSE MIB Index is calculated and published on all days when BIt is open for trading.

FTSE is responsible for the operation of the FTSE MIB Index. FTSE maintains records of the market capitalisation of all constituents and other shares and makes changes to the constituents and their weightings in accordance with the Ground Rules. FTSE carries out reviews and implement the resulting constituent changes as required by the Ground Rules. The FTSE MIB Index constituent shares are selected after analysis of the Italian equity universe, to ensure the Index best represents the Italian equity markets.

The FTSE MIB Index is calculated using a base-weighted aggregate methodology. This means the level of an Index reflects the total float-adjusted market value of all of the constituent stocks relative to a particular base period. The total market value of a company is determined by multiplying the price of its stock by the number of shares in issue (net of treasury shares) after float adjustment. An indexed number is used to represent the result of this calculation in order to make the value easier to work with and track over time. As mentioned above, the Index is computed in real time. The details on how to compute it can be found in [80].

2 Descriptive univariate unconditional statistics

In this section, we separately consider the descriptive univariate unconditional statistics for both the forty assets and for the FTSE MIB Index. By univariate, we mean that, here, we do not consider correlations between the variables under study. By unconditional, we mean that, here, we do not consider the non-stationary and seasonal behavior of the variables under study and the possible memory effects. Correlation and non-stationarity will be discussed in the next section.

2.1 Single Assets

In order to characterize market dynamics on a trade-by-trade level, we consider three variables: the series of time intervals between consecutive trades, τ, the trade volumes, v, and the trade-by-trade logarithmic returns, r. If $p(t_i)$ represents the price of a stock at time t_i where t_i is the epoch of the i-th trade, then we define the return as:

$$r_i = \log \frac{p(t_{i+1})}{p(t_i)} \tag{1}$$

Note that $\tau = t_{i+1} - t_i$ is a random intertrade duration (and not a fixed time interval).

Among the empirical studies on τ, we mention [70, 81], concerning contemporary shares traded over a period of a few months, a study on rarely-traded nineteenth century shares in [82], and results on foreign exchange transactions in [83] and [84].

Tables 2, 3 and 4 contain the descriptive statistics, evaluated for the entire sample, for the time series $\tau_i^h = t_{i+1}^h - t_i^h$ (with $t_0^h = 0$), v_i^h and r_i^h, where the superscript h denotes the specific share.

In Table 2 the third and fourth columns give the two parameters of a Weibull distribution fit. The Weibull distribution has the following survival function:

$$\mathbb{P}(\tau > t) = P(t, \alpha, \beta) = \exp\left[-\alpha t^{\beta}\right] \tag{2}$$

where β is the shape parameter and α is the scale parameter. The values given in Table 2 were fitted using the moment method described in [69]. The quality of these fits is pictorially shown in Fig 1 for A2A, EXO, MS and TIT, respectively. The solid line represents our Weibull fit and the circles are the empirical data. Since different companies have different average intertrade duration $\langle \tau^h \rangle$ (see the second column in Table 2), they are also characterized by a different scale parameter α whereas the shape parameter β is almost the same for all the forty time series. Following [72], a scaling function $P(t^*)$ can be defined:

$$P(t^*) = \exp\left[-(t/\langle \tau \rangle)^{\beta^*}\right] \tag{3}$$

where $\beta^* = \langle \beta \rangle = 0.78$. To test the hypothesis that there is a universal structure in the intertrade time dynamics of different companies, we rescale the survival functions

by plotting them against $t \triangleleft \langle \tau^h \rangle$. We find that, for all companies, data approximately conform to a single scaled plot given by (3) as shown in Fig 2 (see also [69, 72, 85]). Such a behavior is a hallmark of scaling, and is typical of a wide class of physical systems with universal scaling properties [86]. Even if [87] showed that the scaling (3) is far from being universal, at least for the New York Stock Exchange, it is remarkable to find it again for a different index in a different market and seven years later with respect to the findings of [72]. However, to go beyond qualitative estimates, the goodness-of-fit test is given in the sixth column of Table 2, where we report the Anderson-Darling (AD) statistics for the transformed random variable $z_\tau^h = \alpha \tau^\beta$. z_τ should follow an exponential distribution with parameter $\theta = 1$, if τ is distributed according to a Weibull distribution. A glance at Fig 4 immediately shows that this is not the case; for $z_\tau > 4$ there are significant deviations from the exponential law, whereas this is approximately satisfied for $z_\tau \leq 4$. This fact is reflected by the high values of the AD statistics for which the critical value at 0.05 significance level is 1.34. In other words, the Weibull null hypothesis is rejected for all the time series. To confirm these results, we also perform the Lilliefors test. The Lilliefors statistics are larger that the critical value at 0.05 significance level as well. Furthermore, we peform the Kolmogorov-Smirnov test in order to check if the distributions in Fig 4 collapse into one. The results show that the null hypothesis of same distribution is always rejected (contact the authors for full tables). Then, we perform the AD test and the Lilliefors test for the index durations, and also in this case the null hypothesis of Weibull distribution is rejected by both statistical tests. Finally we present results based on the Weibull paper to graphically verify the Weibull distribution hypothesis. As an illustration, Fig 3 shows the Weibull paper for the following assets: A2A, EXO, MS and TIT. We can see the deviation of the empirical data from the straight line expected for the Weibull distribution.

In this paper, we do not study volumes v^h, but we present their descriptive statistics in Table 3 for the sake of completeness.

The descriptive statistics for trade-by-trade returns r^h can be found in Table 4. Notice that there is excess kurtosis. The histograms in Fig 5 for the asset prices A2A, EXO, MS and TIT, respectively, show how the returns are distributed.It is possible to appreciate the discrete character of returns even after the logarithmic transformation.

2.2 FTSE MIB index

As well as the single assets, we investigate the FTSE MIB index. Tables 2 and 4 summarize also the descriptive statistics of the time series τ_i^I and r_i^I respectively evaluated for the FTSE MIB index as trade-by-trade volumes are not available.

In Fig 6 we show the survival function for the intertrade waiting time of the FTSE MIB index. The solid line represents the Weibull fit, whereas the circle represents the empirical data. The shape of the two curves is very different. Therefore, we can immediately see that intertrade times are not Weibull distributed, and, in this case, the fit does not work even as a first approximation. Indeed, for the FTSE MIB index, the standard deviation of intertrade durations is smaller than the average intertrade duration and the AD test and the Lilliefors test reject the null hypothesis of Weibull distribution as discussed previously.

Contrary to the case of single asset returns, the excess kurtosis for the FTSE MIB index is quite large. Fig 7 shows the histogram of the returns for a bin size of 1×10^{-5}.

Following [18], we test the scaling of the empirical returns. As shown in Table 1, the dataset consists of 405560 records for the FTSE MIB index during the period studied (from the 03^{rd} of February 2011 to the 09^{th} of March 2011). From this

database, we compute the new random variable $r^{\Delta}(t;\Delta t)$ defined as:

$$r^{\Delta}(t;\Delta t)=\log\frac{p^{\Delta}(t+\Delta t)}{p^{\Delta}(t)}, \tag{4}$$

where $p^{\Delta}(t)$ is the value of the index at time t. In this way we sample returns on equally spaced and non-overlapping intervals of width Δt. We further assume that the time series is stationary so that it only depends on Δt and not on t (incidentally, we shall see that this is not the case). To characterize quantitatively the experimentally observed process, we first determine the empirical probability density function $P(r^{\Delta}(\Delta t))$ of index variations for different values of Δt. We select Δt equal to 3s, 5s, 10s, 30s and 300s. In Fig 8(a) we present a semi-logarithmic plot of $P(r^{\Delta}(\Delta t))$ for the five different values of Δt indicated above. These empirical distributions are roughly symmetric and are expected to converge to the normal distribution when Δt increases. The null hypothesis of normal distribution has been tested with the Kolmogorov-Smirnov, the Jarque-Bera and the Lilliefors test. The results reported in Table 5 show that the null hypothesis is always rejected.

We also note that the distributions are leptokurtic, that is, they have tails heavier than expected for a normal process. A determination of the parameters characterizing the distributions is difficult especially because larger values of Δt imply a smaller number of data. Again following [18], we study the probability density at zero return $P(r^{\Delta}(\Delta t)=0)$ as function of Δt. This is done in Fig 8(b), where $P(r^{\Delta}(\Delta t)=0)$ versus Δt is shown in a log-log plot. If these data were distributed according to a symmetric α-stable distribution, one would expect the following form for $P(r^{\Delta}(\Delta t)=0)$:

$$P(r^{\Delta}(\Delta t)=0)=\frac{\Gamma(1/\alpha_L)}{\pi\alpha_L(c\Delta t)^{1/\alpha_L}}, \tag{5}$$

where $\Gamma(\cdot)$ is Euler Gamma function, $\alpha_L\in(0,2]$ is the index of the symmetric α-stable distribution and c is a time-scale parameter. The data are well fitted (in the OLS sense) by a straight line of slope $1/\hat{\alpha}_L=0.58$ leading to an estimated exponent $\hat{\alpha}_L=1.72$. The best method to get the values of $P(r^{\Delta}(\Delta t)=0)$ is to determine the slope of the cumulative distribution function in $r^{\Delta}(\Delta t)=0$. In Fig 8(c), we plot the rescaled probability density function according to the following transformation:

$$r^{\Delta}_s=\frac{r^{\Delta}(\Delta t)}{(\Delta t)^{1/\alpha_L}} \tag{6}$$

and

$$P(r^{\Delta}_s)=\frac{P(r^{\Delta}(\Delta t))}{(\Delta t)^{-1/\alpha_L}}, \tag{7}$$

for $\alpha_L=\hat{\alpha}_L=1.72$. Remarkably all the five distributions approximately collapse into a single one. We use the Kolmogorov-Smirnov test to study the null hypothesis of identically distributed rescaled data; the results are shown in Table 6. The null hypothesis is rejected only in the following cases: $\Delta t=3s$ and $\Delta t=5s$, $\Delta t=3s$ and $\Delta t=10s$, $\Delta t=3s$ and $\Delta t=30s$. It is worth noting that this result shows that the scaling, found in the S&P 500 data by Mantegna and Stanley more than twenty years ago [18], still approximately holds in a different market and in a completely different period. We do not run hypothesis tests on the Lévy stable distribution because an eye inspection of Fig 8(c) is sufficient to conclude that the Lévy stable fit is not matching the rescaled data.

3 Descriptive conditional and bivariate statistics

Inspired by [78, 88], in order to study the time variations of the returns during a typical trading day, we use a simple technique. We divide the trading day into equally spaced and non-overlapping intervals of length δt for $\delta t = 3, 5, 10, 30, 300, 600, 900, 1200, 1500$ and 1800 s. Let K be number of intervals and N_k the number of transaction in each interval k. For each interval we evaluate the $\gamma(k)$ indicator as a measure of volatility. $\gamma(k)$ is defined as

$$\gamma(k) = \frac{1}{N_k - 1} \sum_{i=1}^{N_k - 1} \clubsuit_{k,i} - \langle r_k \rangle \clubsuit \tag{8}$$

where $\langle r_k \rangle$ is the average value of returns in the time interval k. In Fig 9(a), as an example, we plot the average value of $\gamma(k)$ over the investigated period as a function of the interval index k for $\delta t = 300$ s. We can see that the volatility is higher in the morning, at the opening of continuous trading, and then it decreases up to midday. There is a local increase after midday and then the volatility returns to lower values to finally grow towards the end of continuous trading. The above pattern can be reinforced by the presence of the many day traders whose practice is to close all their positions at the end of each trading day and reopen them in the following morning. The rationale of day traders is to avoid overnight exposure to risk. Interestingly, this plot also provides us with a picture of the social behavior of Borsa Italiana equity traders. The volatility can be seen to drop off in the interval 12:30 - 14:00 and to grow suddenly again around 14:20. These times correspond to the typically preferred lunchtime interval of most traders. In Fig 9(b), we plot the number of trades on the FTSE MIB index as a function of the interval index k for $\delta t = 300$ s. The behavior of the trade activity closely follows the behavior of volatility. This is even clearer from the analysis of Fig 9(c) where the volatility is plotted as a function of the activity.The scatter plot shows a strong correlation between the two variables. This result does not depend on the length of the interval w, but the corresponding plots are not presented here for the sake of compactness. This feature was already present in the Australian market studied for a much longer period (10 years $\approx$ 2500 days) by [78, 88]. Again, it is remarkable to see a statistical pattern still valid in a different market after more than 10 years.

Fig 9 shows a clear seasonal pattern in intraday trades. In order to take this behavior into account, we proposed to use a non-stationary normal compound Poisson process with volatility of jumps proportional to the activity of the Poisson process in [68]. Here, we take even a more pragmatic stand and we do not assume any a priori relationship between volatility and activity as it emerges spontaneously, if present, with the method described in the next section.

Empirical studies of volatility for daily financial data by [89] have shown that volatility estimates and returns are negatively correlated for positive time lag. Therefore, following [89], we investigate this effect on high frequency data by estimating the leverage correlation function as

$$L(\Delta t) = \frac{\langle (r^2(t + \Delta t))^2 r^2(t) \rangle}{(\mathrm{var}[r^2(t)])^2}, \tag{9}$$

where Δt represents the lag. The estimates for empirical data samples are shown in Fig 10 for $\Delta t = 3$s. The leverage effect is not evident. For comparison, in Figs 11 and 12, we computed $L(\Delta t)$ for 7 major international stock indices (S&P500, NASDAQ, CAC40, FTSE, DAX, Nikkei, Hang Seng) for Δt equal to one day. The dataset consisted of daily close prices adjusted for dividends and splits ranging from January

1990 to October 2000 as in [89]. In the case of S&P500, NASDAQ, DAX, Nikkei and Hang Seng indices, the leverage effect is well evident, whereas for CAC40 and FTSE indices it is less evident. However, in all these cases, the leverage effect is much stronger then in our high frequency data, if any.

4 A compound Poisson type model

As one can see, during a trading day, the volatility and the activity are higher at the opening of the market, then they decrease at midday and they increase again towards market closure [88] (see also Fig 9). In other words, the (log-)price process is non-stationary. As suggested in [68], such a non-stationary process for log-prices can be approximated by a mixture of normal compound Poisson processes (NCPP) in the following way. A normal compound Poisson process is a compound Poisson process with normal jumps. In formula:

$$X(t) = \sum_{i=1}^{N(t)} R_i \, , \tag{10}$$

where R_i are normally distributed independent trade-by-trade log-returns, $N(t)$ is a Poisson process with parameter λ and $X(t)$ is the logarithmic price, $X(t) = \log(P(t))$. By probabilistic arguments one can derive the cumulative distribution function of $X(t)$, it is given by:

$$F_{X(t)}\ (u) = \mathbb{P}(X(t) \leq u) = e^{-\lambda t} \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} F_R^{\star n}(u) \, , \tag{11}$$

where $F_R^{\star n}(u)$ is the n-fold convolution of the normal distribution, namely

$$F_R^{\star n}(u) = \frac{1}{2}\ \ 1 + \mathrm{erf}\ \ \frac{u - n\theta}{\sqrt{2n\sigma^2}}\ \ \, , \tag{12}$$

and θ and σ^2 are the parameters of the normal distribution.

We now assume that the trading day can be divided into n equal intervals of constant activity $\{\lambda_i\}_{i=1}^{n}$ and of length w, then the unconditional waiting time distribution becomes a mixture of exponential distributions and its cumulative distribution function can be written as

$$F_\tau(u) = \mathbb{P}(\tau \leq u) = \sum_{i=1}^{n} a_i (1 - e^{-\lambda_i \tau}) \, , \tag{13}$$

where $\{a_i\}_{i=1}^{n}$ is a set of suitable weights. The activity seasonality can be mimicked by values of λ_i that decrease towards midday and then increase again towards market closure. In order to reproduce the correlation between volatility and activity, one could assume that

$$\sigma_{\xi,i} = c\lambda_i \tag{14}$$

where c is a suitable constant. As already mentioned, however, for practical purposes, one can also estimate three parameters for each interval, the parameter λ_i of the Poisson process and the parameters θ_i and σ_i for the log-returns without any correlation assumptions. This leads us to two possible examples of such compound Poisson type models which will be introduced in Section 4.1 alongside the popular ACD model for later comparisons. After a brief error analysis of the maximum likelihood estimation (MLE) method in Section 4.3 we will move on to the main

Monte Carlo experiment to test model selection using information criteria (IC) in Section 4.4. The different nature of the compound Poisson models and the ACD model makes a direct comparison in terms of model selection questionable. Therefore, our main focus will be a comparison of IC within each model class separately.

4.1 Model definitions and likelihood functions

4.1.1 The compound Poisson model with discrete intensity (D λ)-model

We extend the notation of Eq (10) by an additional index denoting the corresponding interval: We suppose that high-frequency data is given over a time interval $[0,T]$. First, set a time grid $\{t_i\}_{i\in\{1,\dots,n\}}$ such that $t_0 < t_1 < t_2 < \dots < t_n = T$. On each time interval $[t_{i-1},t_i)$ we have a compound Poisson process

$$X_i(t) := \sum_{k=1}^{N_i(t)} R_k^{(i)}, \tag{15}$$

where $\{R_k^{(i)}\}_{k\in\mathbb{N}}$ is an i.i.d. sequence of $\mathcal{N}(\theta_i,\sigma_i^2)$ distributed random variables and $(N_i(t))_{t\geq 0}$ is a homogeneous Poisson process with parameter λ_i. Further, $\{R_k^{(i)}\}_{k\in\mathbb{N}}$ are all independent of $(N_i(t))_{t\geq 0}$.

For a fixed time interval $[t_{i-1},t_i)$ the log-likelihood function is given by

$$\mathcal{L}_i^D(\lambda_i,\theta_i,\sigma_i) = -\lambda_i(t_i - t_{i-1}) + \ln(\lambda_i)N_i(t_i) + \sum_{k=1}^{N_i(t_i)} \ln(p_{\mu_i,\sigma_i}(R_k^{(i)})), \tag{16}$$

where p_{μ_i,σ_i} denotes the probability density function of the $\mathcal{N}(\theta_i,\sigma_i^2)$ distribution. Due to the independence assumptions the overall log-likelihood is given by the sum of all $\mathcal{L}_i$. Eq (16) can be derived from the general expression for the sample density function given on p. 200 in [90] by substituting a constant λ.

The maximum likelihood estimators are therefore:

$$\hat{\lambda}_i = N_i/w_i; \qquad \theta_i = \frac{1}{N_i}\sum_{k=1}^{N_i} r_i; \qquad \hat{\sigma}_i^2 = \frac{1}{N_i}\sum_{k=1}^{N_i}(r_i - \theta_i)^2, \tag{17}$$

where N_i is the number of trades in the ith interval and $w_i = t_i - t_{i-1}$.

Note that the maximum likelihood estimator for σ^2 is biased and the bias can be corrected by using

$$\tilde{\sigma}_i^2 = \frac{1}{N_i - 1}\sum_{k=1}^{N_i}(r_i - \theta_i)^2 \tag{18}$$

instead. We shall use either the biased or unbiased estimator in the following sections when appropriate.

4.2 Approximating stylized facts using the (Dλ)-model

A first Monte Carlo simulation of the (Dλ)-model was performed by considering a trading day divided into a number of intervals of length $w = 3, 5, 10, 30, 300$ s. The parameters $\hat{\lambda}_i$, θ_i and $\tilde{\sigma}_i^2$ were estimated as explained above. Note that we use the unbiased estimator $\tilde{\sigma}_i$ from (18). In the following, we shall focus on estimates based on the FTSE MIB index. Fig 13 displays the histogram of simulated returns and can be compared to Fig 7. In Fig 14, we empirically show that the simulation gives a better fit for the empirical returns of the index as w becomes smaller. This is an encouraging

result meaning that it will be useful to study the convergence of the approximation by means of measure-theoretical probabilistic methods. In order to show that this approximation is able to reproduce the approximate stylized facts described above, Fig 15 shows the scaling relations discussed in section 1.2 for the simulation with $w = 10$ s. The null hypothesis of normal distribution has been tested with the Kolmogorov-Smirnov, the Jarque-Bera and the Lilliefors test. The results reported in Table 7 show that the null hypothesis is always rejected.

One can see from Fig 15(b) that an OLS index estimate $\hat{\alpha}_L = 1.59$ is recovered from the simulation instead of 1.72 for the real index. The scaling given in Eqs. (6), (7) is presented in Fig 15(c), one can see that the approximate scaling still holds for the simulated data. The null hypothesis of identical distribution has been tested with the Kolmogorov-Smirnov test, and the results have been shown in Table 8. It is worth noting that the null hypothesis of identical distribution is always rejected but the statistical value is very near to the critical value.

In Fig 16, we can see that there is no clear leverage effect in the simulated data as in the real case. Finally, in Fig 17, for the simulated time series, we repeat the same analysis presented in Fig 9. Given that, by construction, the non-stationary behavior of the simulated data is modeled on the non-stationary behavior of the real data, it is no surprise to find a qualitative match between the two analyses (see Figs. 9,17).

4.2.1 The compound Poisson model with parametrized intensity (P λ)-model

This model will be used for simulation later on as well as serve as a benchmark model when testing model selection criteria. As empirical results about the trading intensity suggest a daily seasonality, this model assumes that the step function in the (Dλ) model is parametrized by a quadratic function:

$$\lambda_{a,b,c}(t) = at^2 + bt + c \quad t \in [0, 1] \tag{19}$$

Of course, this parametrization can be easily replaced by more complicated functions. Since λ needs to be positive and convex, we also have the conditions

$$a > 0 \text{ and } c > \frac{b^2}{4a} \tag{20}$$

Similar to the (Dλ)-model, the log-likelihood for the (Pλ)-model is given by

$$\mathcal{L}_i^P(a, b, c, \theta_i, \sigma_i) = -\lambda_{a,b,c}(t_{i-1})(t_i - t_{i-1}) + \ln(\lambda_{a,b,c}(t_{i-1}))N_i(t_i) + \sum_{k=1}^{N_i(t_i)} \ln(p_{\mu_i,\sigma_i}(R_k^{(i)})) \tag{21}$$

While the maximum likelihood estimators for θ_i and σ_i are the same as for the (Dλ) case, the maximum likelihood estimators for a, b, c , which determine the form of λ, cannot be obtained in closed form. As a consequence, a numerical optimization method needs to be applied to estimate those parameters.

4.2.2 The ACD model

The autoregressive conditional duration model was first proposed by Engle and Russell [33]. We will consider a model for the durations between events only without marks: Let $(\varepsilon_i)_{i \in \mathbb{N}}$ be a sequence of i.i.d. random variables. The autoregressive conditional duration (ACD) model is defined as

$$x_i = \psi_i \varepsilon_i \tag{22}$$

$$\psi_i \equiv \psi_i(x_{i-1}, \ldots, x_1; \theta) := \mathbb{E}[x_i | x_{i-1}, \ldots, x_1] \tag{23}$$

The innovations (ε_i) are assumed to follow an exponential distribution, i.e. $\varepsilon_i \sim \text{Exp}(1)$, and ψ_i has the following representation

$$\psi_i := \omega + \sum_{j=0}^{m} \alpha_j \, x_{i-j} \; + \sum_{j=0}^{q} \beta_j \, \psi_{i-j} \,, \tag{24}$$

where $\omega > 0$, $\alpha_i \geq 0$ and $\beta_i \geq 0$ for all i. We will call this model ACD(m,q). For given duration data $(x_1, \ldots, x_n)$ the log-likelihood function is given by

$$\mathcal{L}^{\text{ACD}}(\omega, \alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_q) = -\sum_{i=1}^{n} \ln \psi_i + \frac{x_i}{\psi_i} \tag{25}$$

(see p. 104 in [20]).

4.3 MLE and goodness of fit

Before we turn our attention to the actual model selection procedure, it is useful to get a rough idea about how well the underlying MLE method works for the three model classes. We would also like to ensure that the MLE method works reasonably well since a poor ML fit might compromize the quality of the order selection. Due to asymptotic results, we expect that goodness of fit and correctness of the model selection procedure should improve with increasing size of the underlying sample. As these two effects are closely related, it is hard to quantify them separately.

In the next sections, we give a detailed explanation on the simulation procedure and on how the parameter estimation is implemented. Based on that, we run a MLE on previously generated mock data. As we know the true parameter values, we can easily calculate the mean squared error (MSE) as measure for the goodness of fit.

4.3.1 Compound Poisson models

Simulation The simulation algorithm essentially uses the (Pλ)-model. For simplicity we will choose the time interval $[t_0, T]$ to be $[0, 1]$. For the simulation we set an equidistant grid $0 = t_0 < t_1 < t_2 < \ldots < t_n = 1$ on the time interval. Thus, the interval $[0, 1]$ is divided into n subintervals. For $i \in \{1, \ldots, n\}$ the parameters θ_i, σ_i and λ_i on the subinterval $[t_{i-1}, t_i)$ are chosen to be

$$\theta_i = 0, \quad \sigma_i = 1 \quad \text{and} \quad \lambda_i = \lambda(t_{i-1}) \; \forall i \in \{1, \ldots, n\}, \text{ where} \tag{26}$$

$$\lambda(t) := 4(\lambda_{\max} - \lambda_{\min})(t - 0.5)^2 + \lambda_{\min}, \; \forall t \in [0, 1] \text{ and } \lambda_{\min}, \lambda_{\max} > 0 \text{ constant}. \tag{27}$$

The functional form of λ is inspired by the empirical findings in the previous sections and should account for the observed seasonality in a simple way. Of course, the functional form of λ can be easily replaced by more complex functions. We have chosen $\lambda_{\min} = 100$ and $\lambda_{\max} = 10000$. Note that the $\{\lambda_i\}$ form a step function approximation of the parabola in (27). For different grid sizes, we simulate with sample size 1000 each.

Fitting The fitting will be carried out using different grid sizes. Note that the grid size to be used in fitting is bounded from above by the length of the entire time interval (in our case 1). However, as we would like to emulate the behavior of the intensity which was observed in empirical data, i.e. high intensity at the beginning and at the end of the trading day and relatively low intensity in the middle of the day. Consequently, we need at least 3 subintervals to have a piecewise constant function that fulfils these conditions on the time interval. Further, the smallest eligible grid size

is bounded from below by the maximal distance between neighbouring data points within the data set. Otherwise, there are subintervals which do not contain any data points. In such cases, the estimation formulas in (17) would fail.

More precisely, for the maximal distance Δ_{max} between two consecutive data points within a given sample, the finest valid equidistant grid has at most $\left\lfloor \frac{1}{\Delta_{max}} \right\rfloor$ subintervals. Therefore, we will consider a list of candidate models on grids which correspond to $n = 3, 4, \ldots, \left\lfloor \frac{1}{\Delta_{max}} \right\rfloor$ subintervals on the interval $[0, 1]$.

For the (Dλ) model, the estimators are given in closed form in (17) and the likelihood value is easily calculated via Eq (16) and subsequently used for the calculation of the IC. We decide to use the unbiased estimator $\hat{\sigma}_i^2$: Since we are mainly interested in model selection, we would like to ensure that we work with the optimal value of the log-likelihood when calculating the IC (see also 4.4).

In order to fit the (Pλ) model, we assume that the estimates for $\rho\theta_i\rangle$, $\rho\sigma_i\rangle$ and $\rho\lambda_i\rangle$ for the (Dλ)-algorithm are already calculated and can be used as an input for the estimation of the (Pλ)-model. As mentioned previously, the estimators for θ_i and σ_i coincide in both models and no further calculation is needed for these parameters. It remains to solve the following minimization problem:

$$(\hat{a}, \hat{b}, \hat{c}) = \underset{a,b,c \in R}{\arg\min} \left[-\sum_{i=1}^{n} \mathcal{L}_i^P \left(a, b, c, \theta_i, \sigma_i \right) \right] \quad \text{s.t.} \quad a > 0 \text{ and } c > \frac{b^2}{4a} \tag{28}$$

A reasonable choice of the starting value for the minimization algorithm can be easily obtained by the least-squares fit of the parabola to the $\rho\lambda_i\rangle$ values of the (Dλ) case, which already gives a fairly good approximation of the parabola. In case the initial values obtained by this method do not lie in the admissible set, a change of signs for a or a shift of the parabola may be applied.

Note that the estimation of the (Pλ)-model requires a grid with at least 4 grid points, i.e. 3 subintervals on which $\lambda_1, \lambda_2, \lambda_3$ are estimated using the (Dλ)-model. This ensures that the parabola is well determined. However, as mentioned before, this condition is not restrictive and covers all models on which we would like to run model selection.

4.3.2 ACD model

For both simulation and MLE of ACD models we use the R package ACDm written by Markus Belfrage [91]. The model selection analysis for the ACD model follows the Monte Carlo experiment conducted in [92]. We consider model orders $m, q \in \{1, 2\}$ and Table 9 shows the choice of parameters for the simulation.

4.3.3 Numerical results

We use the MSE as a measure for the goodness of fit: Let θ be a generic model parameter to be estimated and $\hat{\theta}$ the corresponding estimator. Given $N = 1000$ samples and $\hat{\theta}^{(k)}$, $k = 1, \ldots, N$, the estimates for each sample we calculate the mean squared error to be

$$\mathrm{MSE}(\theta) = \mathbb{E}\left[\left(\theta - \hat{\theta} \right)^2 \right] = \frac{1}{N} \sum_{k=1}^{N} \left(\theta - \hat{\theta}^{(k)} \right)^2 \tag{29}$$

Compound Poisson models We have to point out first that the distance in Eq (29) has to be understood as a functional distance. To be more precise, we choose the

L^2-distance between the true step function intensity and the estimated one:

$$\mathbb{E}\left[\clubsuit - \hat{\theta}\clubsuit\right] = \mathbb{E}\left[\|\theta - \hat{\theta}\|_{L^2}^2\right] \tag{30}$$

The cases of θ and σ^2 are the easier ones, as we just need to calculate the distance between a step function and a constant: For the step functions with values $\P\theta_i\diamondsuit$on the fitting grid $t_1 < t_2 < \triangleright\triangleright\triangleright < t_n$ Eq (30) can be further written as

$$\mathbb{E}\ \|\theta - \theta\|_{L^2}^2 = \frac{1}{N}\sum_{k=1}^{N}\|\theta - \theta^{(k)}\|_{L^2}^2 = \frac{1}{N}\sum_{k=1}^{N}\int_0^T (\theta(t) - \theta^{(k)}(t))^2 dt \tag{31}$$

$$= \frac{1}{N}\sum_{k=1}^{N}\sum_{i=2}^{n}(\theta - \theta_i^{(k)})^2(t_i - t_{i-1}) \tag{32}$$

$$\tag{33}$$

and in the same way for σ^2.

Concerning the intensity function, we have to merge the simulation grid $t_1^s < t_2^s < \triangleright\triangleright\triangleright < t_m^s$ with the fitting grid $t_1^f < t_2^f < \triangleright\triangleright\triangleright < t_r^f$. After reordering and relabeling, we can calculate the MSE on the merged grid $t_1 < t_2 < \triangleright\triangleright\triangleright < t_n$ via

$$\mathbb{E}\left[\|\lambda - \hat{\lambda}\|_{L^2}^2\right] = \frac{1}{N}\sum_{k=1}^{N}\sum_{i=2}^{n}(\lambda_i - \hat{\lambda}_i^{(k)})^2(t_i - t_{i-1})\triangleright \tag{34}$$

The numerical results we present here are exemplarily for $N = 1000$ samples of data simulated from a grid containing 30 subintervals.

Table 10 shows summary statistics of θ and σ^2, where the summary statistics were calculated over the set of fitting grids. The MSE for the θ and σ^2 are comparably small.

For the intensity function λ we plot the MSE against the number of subintervals used for fitting in Fig 18. Starting from a small number of subintervals, the MSE decreases sharply before it reaches its optimum at 30, the true number of subintervals from the simulation. Number of subintervals above 30 give a larger MSE and, in the case of the (Dλ) model, instabilities of over parametrization even lead to an increasing MSE.

Concerning goodness of fit, we can see that the MSE of the (Pλ) model is consistently smaller than the MSE of the (Dλ) model. This is to be expected as by construction of the experiment the (Pλ) model is the true model and gives a better fit to the data.

Moreover, we can observe that apart from the optimum at 30 there are "preferred" numbers of subintervals at 10, 20, 45, 60. This is crucial for the explanation of the behavior of model selection as the relationship between goodness of fit and number of subintervals in the region below the optimal number is not monotone.

One might be concerned about the large values of the MSE of the λ estimation. However, first note that the sample size is neither controlled by the choice of simulation grid size nor the fitting grid size. The sample size is determined by the value of the intensity λ. Consequently, if the fitting grid is already sufficiently fine, the sample size is approximately of the same order. Since, the sample size does not change much for finer fitting grids, we therefore cannot expect to observe any convergence of the MSE to 0 in Fig 18.

Second and more importantly, the size of the MSE can be estimated by the expected fluctuations of the estimator $\hat{\lambda}$. The MSE can be estimated from below by the ideal situation when the simulation and fitting grid are identical. Without loss of

generality, we assume an equidistant simulation grid with grid size $w = t_i - t_{i-1}$ and rewrite Eq (34):

$$\mathbb{E}\left[\|\lambda - \hat{\lambda}\|^2_{L^2}\right] \geq w \sum_{i=2}^{n} \mathbb{E}\left[(\lambda_i - \hat{\lambda}_i)^2\right] = w \sum_{i=2}^{n} \mathrm{Var}\left[\hat{\lambda}_i\right] = \frac{1}{w} \sum_{i=2}^{n} \mathrm{Var}\left[N_i\right], \tag{35}$$

where we have used the definition of the estimator in (17) and that the number of events in an interval of size w is Poisson distributed: $N_i \sim \mathrm{Poi}(\lambda w)$. We finally get that

$$\mathbb{E}\left[\|\lambda - \hat{\lambda}\|^2_{L^2}\right] \geq \frac{1}{w} \sum_{i=2}^{n} \mathrm{Var}\left[N_i\right] = \frac{1}{w} \sum_{i=2}^{n} \lambda_i w \approx \frac{1}{w} \int_0^1 \lambda(t)\, \mathrm{d}t, \tag{36}$$

where we approximate the integral of the step function by the integral of the smooth intensity parametrization in Eq (27). For our numerical example we have $\frac{1}{w} = 30$ and $\lambda_{\min} = 100$ and $\lambda_{\max} = 10000$. An explicit calculation of above integral gives the rough estimate

$$\mathbb{E}\left[\|\lambda - \hat{\lambda}\|^2_{L^2}\right] \gtrsim 30 \leq 3400 = \mathcal{O}(10^5), \tag{37}$$

which is of about the same order of magnitude observable in Fig 18.

ACD model In the ACD case we have a simple parameter vector $(\omega, \alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_q) \in \mathbb{R}^{1+m+q}$, Therefore, we can use the formula given in Eq (29) for each scalar valued parameter. The results can be seen in Table 11. The largest sample size ensures that the MSE are comparably low for each model. The largest contribution to the MSE comes from the ω parameter. An even closer look shows that the MSE of the β parameter(s) is of different order depending on the model order q. In the case $q = 1$, the MSE of the β parameter is of the same size as the α parameter(s). However, in the case of $q = 2$, the order of the MSE of the β parameters are significantly larger than the MSE of the α parameters (by a factor of 10 in the ACD(1, 2) case and by a factor of 100 in the ACD(2, 2) case).

4.4 Information criteria and model selection

Starting off from the estimation results in the previous section, we would like to analyse how effective model selection based on information criteria (IC) performs for both the coumpounds Poisson models and the ACD model.

As seen in the previous Monte Carlo simulation choosing smaller values of w, i.e. increasing the number of model parameters, gives better fits and the model is able to capture all distributional properties of the quantity of interest. However, a model containing a large number of parameters is likely to be over fitted. A quantitative method to resolve this trade-off situation is to apply IC. In the following, we will consider three of the most common information criteria:

For a given model fitted to data via MLE let $\mathcal{L}$ be the maximal log-likelihood value, k the number of parameters and T be the sample size of the data set. Then we define:

1. **Akaike's information criterion (AIC)** (see [93])

$$\mathrm{AIC} = -2\mathcal{L} + 2k \tag{38}$$

2. **Bayesian information criterion (BIC)** (see [94])

$$\mathrm{BIC} = -2\mathcal{L} + k \ln(T) \tag{39}$$

3. Hannan and Quinn information criterion (HQ) (see [95] and [96])

$$\mathrm{HQ} = -2\mathcal{L} + 2k\ln(\ln(T)) \tag{40}$$

Note that the information criteria under consideration penalize the log-likelihood value for increasing number of parameters k. Among several candidate models, one chooses the model with the smallest IC value. A time grid $t_0 < t_1 < \triangleright\triangleright\triangleright < t_n$ is given and divides the overall time interval in n subintervals. Recall that we from Section 4.3.1 that we do not consider $n \in \P 1, 2\diamondsuit$. Then the (D$\lambda$)-model has in total $k = 3n$ parameters with $n \in \P 3, 4, \triangleright\triangleright\triangleright\diamondsuit$. This will also be the true number of parameters we expect the IC to choose. In the same way we have for the (Pλ)-model $k = 2n + 3$ parameters with $n \in \P 3, 4, \triangleright\triangleright\triangleright\diamondsuit$.

4.4.1 Numerical results

Compound Poisson models Figs 19, 20 and 21 show box plots of the model selection results of the AIC, BIC and HQ respectively. In each box plot, the orange and blue box plot correspond to the results of the (Dλ)- and (Pλ)-model respectively. The horizontal axis shows the number of subintervals used in the simulation grid. On the vertical axis are the selected number of parameters after the parameter estimation of the (Dλ)- and (Pλ)-models using different discretizations of $[0, 1]$. A single box in the box plots extends from the 25th percentile to the 75th percentile and the dot indicates the median. The whiskers have a maximum length of 1.5 times the box length and extend to the outermost point which is not considered as outlier. The crosses indicate outliers.

Below the box plots, bars indicate the ratio of samples which allow model selection under correct specification (blue) and under misspecification (red): In our setting, we speak of model selection under misspecification if the correct model is not contained in the set of selectable models and cannot be chosen by the IC. If this is not the case, i.e. the correct model can potentially be chosen by the IC, we call it model selection under correct specification.

The results for the (Dλ) and (Pλ) model are very similar. Common for all three IC is that for small parameter numbers below 15 the model selection works well: the distributions of the selected orders are concentrated and closely follow the $3n$ or $2n + 3$ reference line respectively, where n is the number of subintervals. For very large parameter numbers one can observe that the selected model orders remain distributed around a maximum model order and stop to follow the linear trend of the reference line. This is rather due to the limitations of our MC setup than the inherent property of the IC: As described in Section 4.3.1, we only work with equidistant grids when applying the model selection procedure. The finest grid which can be used for fitting is determined by the maximal distance Δ_{max} between two consecutive points within a sample. On the other hand, Δ_{max} is related to the minimal value of λ in the middle of the interval., depending on how small we choose the simulation grid size Δ_{sim}. This means that whenever $\Delta_{\mathsf{max}} > \Delta_{\mathsf{sim}}$, the true model is not contained in the pool of models from which the IC may choose from. In other words, we have a case of model selection under misspecification. The bar plots show that first cases occur at around $n = 20$ and go up to a ratio of about 50% for the finest grid in the analysis.

Another look at Fig 18 hints that the general rule "the more parameters, the better the fit" is not entirely true: we can observe that the relation between grid size and MSE is not entirely monotone. This is due to the fact that the fit of the specific model does not only depend on the number of parameters, but also to some extent on the position of the grid. As a consequence, under misspecification, the selected order does not necessarily correspond to the finest available grid size above Δ_{sim}. This might explain the "plateaus" on the model selection results for large parameters.

Between the region of very small and very large parameters the IC exhibit quite different behaviors according to their intrinsic tendency of under- and overfitting, which will be described in the following:

The AIC tends to overestimate the number of parameters. It allows outliers (in the region of $n \leq 22$) as well as a larger number of cases of the model selection to lie above the reference line (in the region of $n \geq 23$). In contrast, the selected model orders of the BIC and HQ are either on the reference line or strictly below the reference line. In other words BIC and HQ tend to underestimate. Additionally, we can see that for the AIC the boxplot starts to deviate from the reference line starting around $n = 25$ to $n = 27$ and the BIC and HQ deviate earlier around $n = 15$ and $n = 20$ respectively. Especially, for $n < 27$ the underestimation in the BIC and HQ case is not attributable to the behaviour of model selection under misspecification, as the ratio of model selection under misspecification is rather low. Based on our results, If the IC were to be ordered by their parsimonious character, the BIC would be the more parsimonious whereas the AIC the least.

The above observations show that the model selection using any of the three IC works quite well as long as the true model is actually retrievable. The AIC tends to overestimate, but the model selection results are closest to the reference line of true parameters compared to the other two IC.

ACD model The results of the model selection experiment can be found in Tables 12 to 15. The numbers are success rates in percent of the respective IC to select the correct model from which the simulation data was generated from. The qualitative behaviour of the IC are not surprisingly similar to the findings for the GARCH model in [92].

A closer look at Table 12 shows that the success rate of the IC is exceptionally good in the case of ACD(1, 1) data. Even for a small sample size all information criteria are able to detect the correct model order in the majority of cases. The tendency to under fit works in favour for the BIC and to some extent also for the HQ. For the same reason, the success rates for the AIC are relatively low due to its overfitting property.

A similar behaviour can be observed for ACD(2, 1) in Table 14: Although the IC underestimate the model for smaller sample sizes as a ACD(1, 1) model, they improve for large sample sizes.

In both the ACD(1, 1) and the ACD(2, 1) case, i.e. the cases for $q = 1$, the behaviour of the model selection is acceptable: a reasonably large sample size, which is of the order of a typical intra day trading data sample, ensures a sufficiently large success rate in detecting the correct model. Unfortunately, this cannot be said about the case $q = 2$:

In the first example of ACD(1, 2) data in Table 13, we see that the correct model order is never detected in the majority of cases even for large sample sizes. The best success rates are the ones of the AIC again due to its overfitting tendency. This may be concerning, as this shows that despite the fact that ACD(1, 2) and ACD(2, 1) have the same number of parameters the model selection behaviour is far from comparable.

In comparison, the results for the ACD(2, 2), the most complex model in our experiment, are even more critical: Not only are the IC unable to detect the correct model in most of the cases even with large samples, but the best success rates, again from the AIC, are below 20%.

As mentioned in Section 4.3.3, the cases where model selection fails align with relatively high MSE of the β parameters for $q = 2$: The contribution of the MSE of the ω parameter is not as important, as this parameter is included in all models. However, the increase in MSE when moving from $q = 1$ to $q = 2$ might be one of the factors explaining the discrepancy in model selection between $q = 1$ and $q = 2$. This

part of our MC experiment suggests that parameters which are harder to estimate compared to other model parameters (in our case α vs. β parameters or in other words moving average vs. autoregressive parameters in Eq (24), might also be less likely to be detected by model selection.

5 Conclusions

In this paper, we addressed two questions. The first one concerns to so-called stylized facts for high-frequency financial data. In particular, do the statistical regularities detected in the past still hold? We cannot give a negative answer to this question. Indeed, we find that some of the scaling properties for financial returns are still approximately satisfied. Most of the studies we refer to concerned a different market (the US NYSE) and were performed several years ago. However, one of the first econophysics papers (if not the first one) concerned returns in the Italian stock exchange (see [97]) and, for this reason, we decided to focus on this market.

The second question is: Is it possible to approximate the non-stationary behavior of intra-day tick-by-tick returns by means of a simple phenomenological stochastic process? We cannot give a negative answer to this question, so far. In Section 4, we present a simple non-homogeneous normal compound Poisson process and we argue that it can approximate empirical data. The cost for simplicity is potential over-fitting as we have to estimate many parameters, but the outcome is a rather accurate representation of the real process. Whether it is possible to rigorously prove convergence of the method outlined in Section 4 is subject to further research and it is outside the scope of the present paper. It is well-known that Lµevy processes, namely stochastic processes with stationary and independent increments, can be approximated by compound Poisson processes. The method described in Section 4 can provide a clue for a generalization of such a result to processes with non-stationary and non-independent increments.

Concerning the issue of overfitting, the second part of Section 4 shows that IC are able to detect model orders correctly to some extent when applied to simulated data. It remains to check how well the model selection method performs on empirical data. As a consequence from the numerical results, due to the high variability of model selection in the region of larger numbers of parameters it is not advisable to rely only on the IC based model selection. It is recommended to combine these with further cross-validation techniques. A similar conclusion holds for the ACD model, as model selection using IC is adversely affected by differing MLE quality for different model orders.

Acknowledgments

This work was partially supported by MIUR PRIN 2009 grant The growth of firms and countries: distributional properties and economic determinants - Finitary and non-finitary probabilistic models in economics 2009H8WPXY_002 and by an SDF fund provided by the University of Sussex.

References

1. Goodhart C, O'Hara M. High-frequency data in financial markets: Issues and applications. J of Empir Financ. 1997;4:73–114.

2. O′Hara M. Making market microstructure matter. Financ Manage. 1999;28:83–90.

3. Madhavan A. Market microstructure: A survey. Journal of Financial Markets. 2000;3:205–258.

4. Scalas E, Gorenflo R, Mainardi F. Fractional calculus and continuous-time finance. Physica A. 2000;284:376–384.

5. Mainardi F, Raberto M, Gorenflo R, Scalas E. Fractional calculus and continuous-time finance II: the waiting-time distribution. Physica A. 2000;287(3-4):468–481.

6. Dacorogna M, Gen˘cay R, Müller U, Olsen RB, Pictet O. An Introduction to High Frequency Finance. Academic Press; 2001.

7. Raberto M, Cincotti S, Focardi SM, Marchesi M. Agent-based simulation of a financial market. Physica A. 2001;219:319–327.

8. Cincotti S, Focardi SM, Marchesi M, Raberto M. Who wins? Study of long-run trader survival in an artificial stock market. Physica A. 2003;324(1-2):227–233.

9. Luckock H. A steady-state model of the continuous double auction. Quant Finance. 2003;3:385–404.

10. Scalas E, Gorenflo R, Lucklock H, Mainardi F, Mantelli M, Raberto M. Anomalous waiting times in high-frequency financial data. Quant Finance. 2004;4:1–8.

11. Pastore S, Ponta L, Cincotti S. Heterogeneous information-based artificial stock market. New J Phys. 2010;12:053035.

12. Ponta L, Pastore S, Cincotti S. Information-based multi-assets artificial stock market with heterogeneous agents. Nonlinear Anal Real World Appl. 2011;12:1235–1242.

13. Ponta L, Raberto M, Cincotti S. A multi-assets artificial stock market with zero-intelligence traders. Europhys Lett. 2011;93:28002.

14. Ponta L, Scalas E, Raberto M, Cincotti S. Statistical Analysis and Agent-Based Microstructure Modeling of High-Frequency Financial Trading. IEEE Journal of Selected Topics in Signal Processing. 2012;6:381–387.

15. Mandelbrot B. The Variation of certain speculative prices. J Business. 1963;36:394–419.

16. Mandelbrot B. Fractals and Scaling in Finance. Berlin: Springer; 1997.

17. Müller U, Dacorogna M, Olsen RB, Pictet OV, Schwarz M, Morgenegg C. Statistical study of foreign exchange rates. J Bank Financ. 1990;14:1189–1208.

18. Mantegna RN, Stanley HE. Scaling behavior in the dynamics of an economic index. Nature. 1995;376(6535):46–49.

19. Gopikrishnan P, Plerou V, Gabaix X, Stanley HE. Statistical properties of share volume traded in financial markets. Phys Rev E. 2000;62:4493–4496.

20. Hautsch N, editor. Econometrics of Financial High-Frequency Data. Berlin: Springer; 2012.

21. Gontis V, Kaulakys B. Long-range memory model of trading activity and volatility. Journal of Statistical Mechanics. 2006;P10016:1–11.

22. Gontis V, Kaulakys B. Modeling long-range memory trading activity by stochastic differential equations. Physica A. 2007;382:114–120.

23. Gontis V, Ruseckas J, Kononovicius A. A long-range memory stochastic model of the return in financial markets. Physica A. 2010;389:100–106.

24. Gontis V, Kononovicius A. Nonlinear Stochastic Model of Return matching to the data of New York and Vilnius Stock Exchanges. Dynamics of Socio-Economic Systems. 2011;2:101–109.

25. Kaulakys B, Alaburda M, Gontis V. Point Processes Modeling of time series exhibiting power-law statistics. AIP Conference Proceedings. 2007;922:535– 538.

26. Kaulakys B, Alaburda M, Gontis V, Meskauskas T, Ruseckas J. Modeling of Flows with Power-law Spectral Densities and Power-law Distributions of Flow Intensities. In: Schadschneider A, editor. Traffic and granular flow. vol. 5. Springer; 2007. p. 587–594.

27. Kaulakys B, Alaburda M, Gontis V. Long-range stochastic point processes with the power law statistics. In: M Janzura MH, editor. Proceeding of Prague Conference. Charles University in Prague, Prague: Matfyzpress; 2006. p. 364–373.

28. Kaulakys B, Alaburda M, Gontis V, Meskauskas T. Multifractality of the Multiplicative Autogressive Point Processes. In: Novak MM, editor. Complexus Mundi: Emergent Patterns in Nature. World Scientific; 2006. p. 277–286.

29. Kenett DY, Ben-Jacob E, Stanley HE, Gur-Gershgoren G. How High Frequency Trading Affects a Market Index. Scientific Reports. 2013;3:2110.

30. Zheng Z, Qiao Z, Takaishi T, Stanley HE, Li B. Realized volatility and absolute return volatility: A comparison indicating market risk. PLOS ONE. 2014;9(7):e102940.

31. Botta F, Moat HS, Stanley HE, Preis T. Quantifying stock return distributions in financial markets. PLOS ONE. 2015;10(9):e0135600.

32. Engle RF, Russell JR. Forecasting the frequency of changes in quoted foreign exchange prices with the autoregressive conditional duration model. J Empir Financ. 1997;4:187–212.

33. Engle RF, Russell JR. Autoregressive conditional duration: A new model for irregularly spaced transaction data. Econometrica. 1998;66:1127–1162.

34. Bauwens L, Giot P. The logarithmic ACD model: An application to the bid-ask quote process of three NYSE stocks. Ann Econ Stat. 2000;60:117–149.

35. Lo AW, MacKinlay AC, Zhang J. Econometric models of limit-order executions. J Financ Econ. 2002;65(1):31–71.

36. Cont R, Bouchaud JP. Herd behavior and aggregate fluctuations in financial markets. Macroecon Dyn. 2000;4(2):170–196.

37. Chowdhury D, Stauffer D. A generalized spin model of financial markets. Eur Phys J B. 1999;8:477.

38. Hardle W, Kirman A. Neoclassical demand - A model-free examination of price-quantity relations in the Marseilles fish market. J Econometrics. 1995;67:227–257.

39. Levy M, Levy H, Solomon S. Microscopic Simulation of the Stock Market: The effect of microscopic diversity. J Phys I France. 1995;5:1087–1107.

40. Lux T, Marchesi M. Scaling and criticality in a stochastic multi-agent model of a financial market. Nature. 1999;397(6718):498–500.

41. Stauffer D, Sornette D. Self-Organized Percolation Model for Stock Market Fluctuations. Physica A. 1999;271:496–506.

42. Youssefmir M, Huberman BA. Clustered volatility in multiagent dynamics. J Econ Behav Organ. 1997;32(1):101–118.

43. Hawkes A. Spectra of some self-exciting and mutually exciting point processes. Biometrika. 1971;58:83–90.

44. Muni Toke I, Pomponio F. Modelling Trades-Through in a Limit Order Book Using Hawkes Processes. Economics: The Open-Access, Open-Assessment E-Journal. 2012;6:2012–22.

45. Filimonov V, Sornette D. Apparent criticality and calibration issues in the Hawkes self-excited point process model: application to high-frequency financial data. Quant Finance. 2015;15:1293–1314.

46. Blume ME, Mackinlay AC, Terker B. Order Imbalances and Stock Price Movements on October 19 and 20, 1987. J Finance. 1989;44:827–848.

47. Chan K, Fong WM. Trade size, order imbalance, and the volatility-volume relation. J Financ Econ. 2000;57:247–273.

48. Stoll HR, Whaley RE. Stock Market Structure and Volatility. Rev Financ Stud. 1990;3:37–71.

49. Hauser S, Lauterbach B. The Impact of Minimum Trading Units on Stock Value and Price Volatility. J Financ Quant Anal. 2003;38:575–589.

50. Chordia T, Roll R, Subrahmanyam A. Order imbalance, liquidity, and market returns. J Financ Econ. 2002;65:111–130.

51. Ponzi A, Lillo F, Mantegna RN. Market reaction to a bid-ask spread change: A power-law relaxation dynamics. Phys Rev E. 2009;80:016112.

52. Svorencik A, Slanina F. Interacting gaps model, dynamics of order book, and stock-market fluctuations. Eur Phys J B. 2007;57:453–462.

53. Wyart M, Bouchaud JP, Kockelkoren J, Potters M, Vettorazzo M. Relation between bid-ask spread, impact and volatility in order-driven markets. Quant Finance. 2008;8:41–57.

54. Moro E, Vicente J, Moyano LG, Gerig A, Doyne Farmer J, Vaglica G, et al. Market impact and trading profile of hidden orders in stock markets. Phys Rev E. 2009;80:066102.

55. Perelljo J, Masoliver J, Kasprzak A, Kutner R. Model for interevent times with long tails and multifractality in human communications: An application to financial trading. Phys Rev E. 2008;78:036108.

56. Preis T, Schneider JJ, Stanley HE. Switching processes in financial markets. Proc Natl Acad Sci USA. 2011;108:7674–7678.

57. Kumaresan M, Krejic N. A model for optimal execution of atomic orders. Comput Optim Appl. 2010;46:369–389.

58. Zaccaria A, Cristelli M, Alfi V, Ciulla F, Pietronero L. Asymmetric statistics of order books: The role of discreteness and evidence for strategic order placement. Phys Rev E. 2010;81:066101.

59. Lim M, Coggins R. The immediate price impact of trades on the Australian Stock Exchange. Quant Finance. 2005;5:365–377.

60. Weber P, Rosenow B. Order book approach to price impact. Quant Finance. 2005;5:357–364.

61. Bouchaud JP. The subtle nature of financial random walks. Chaos. 2005;15:026104.

62. Bonanno G, Lillo F, Mantegna R. Dynamics of the number of trades of financial securities. Physica A. 2000;280:136–141.

63. Plerou V, Gopikrishnan P, Amaral LAN, Gabaix X, Stanley HE. Economic fluctuations and anomalous diffusion. Phys Rev E. 2000;62:3023–3026.

64. Scalas E. The application of continuous-time random walks in finance and economics. Physica A. 2006;362:225–239.

65. Gontis V, Kaulakys B, Ruseckas J. Trading activity as driven Poisson process: Comparison with empirical data. Physica A. 2008;387:3891–3896.

66. Hudson RL, Mandelbrot BB. The (Mis)Behaviour of Markets. Profile Business; 2010.

67. Vrobel S. Fractal Time Why a Watched Kettle Never Boils. In: West BJ, editor. Studies Of Nonlinear Phenomena In Life Science. Imperial College Press: World Scientific; 2011.

68. Scalas E. Mixtures of compound Poisson processes as models of tick-by-tick financial data. Chaos Soliton Frac. 2007;34:33–40.

69. Politi M, Scalas E. Fitting the empirical distribution of intertrade durations. Physica A. 2008;387:2025–2034.

70. Raberto M, Scalas E, Mainardi F. Waiting-times and returns in high-frequency financial data: an empirical study. Physica A. 2002;314(1-4):749–755.

71. Masoliver J, Montero M, Weiss GH. Continuous-time random-walk model for financial distributions. Phys Rev E. 2003;67:021112.

72. Ivanov PC, Yuenand A, Podobnik B, Lee Y. Common scaling patterns in intertrade times of U.S. stocks. Phys Rev E. 2004;69:056107.

73. Goldstein ML, Morris SA, Yen GG. Problems with fitting to the power-law distribution. Eur Phys J B. 2004;41:255–258.

74. Embrechts P, Klüppelberg C, Mikosch T. Modelling Extremal Events for Insurance and Finance. Berlin: Springer; 1997.

75. Press WH, Flannery BP, Teukolsky SA. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press; 1992.

76. Mainardi F, Gorenflo R, Scalas E. A fractional generalization of the Poisson process. Vietnam J Math. 2004;32:53–64.

77. Engle RF, Russell JR. Forecasting transaction rates: the autoregressive conditional duration model. NBER Working paper series. 1994; p. 4966.

78. Bertram WK. A threshold model for Australian Stock Exchange equities. Physica A. 2005;346:561–576.

79. Livan G, Inoue J, Scalas E. On the non-stationarity of financial time series: Impact on optimal portfolio selection. J Stat Mech. 2012; p. P07025.

80. FTSEMIB. Methodology for the management of the FTSE MIB index. London, U.K.: Borsa Italiana, London Stock Exchange Group; 2011.

81. Golia S. Long memory effects in ultra- high frequency data. Quaderni di Statistica. 2001;3:43–52.

82. Sabatelli L, Keating S, Dudley J, Richmond P. Waiting time distributions in financial markets. Eur Phys J B. 2002;27:273–275.

83. Takayasu H, editor. Empirical Science of Financial Fluctuations: The Advent of Econophysics. Tokyo: Springer; 2002.

84. Marinelli C, Rachev ST, Roll R. Subordinated exchange rate models: Evidence for heavy tailed distributions and long-range dependence. Math Comput Modell. 2001;34:955–1001.

85. Stauffer D, Stanley HE. From Newton to Mandelbrot: A Primer in Theoretical Physics. Berlin: Springer; 1995.

86. Bunde A, Havlin S, editors. Fractals in Science. Berlin: Springer; 1994.

87. Eisler Z, Kertjesz J. Size matters: some stylized facts of the stock market revisited. Eur Phys J B. 2006;.

88. Bertram WK. An empirical investigation of Australian Stock Exchange data. Physica A. 2004;341:533–546.

89. Bouchaud JP, Matacz A, Potters M. Leverage Effect in Financial Markets: The Retarded Volatility Model. Phys Rev Lett. 2001;87:228701.

90. Snyder DL, Miller MI. 4.4. In: Random Point Processes in Time and Space. Springer; 1991.

91. Belfrage M. ACDm: Tools for Autoregressive Conditional Duration Models; 2016. Available from: https://CRAN.R-project.org/package=ACDm

92. Javed F, Mantalos P. GARCH-type models and performance of information criteria. Comm Statist Simulation Comput. 2013;42(8):1917–1933.

93. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory (Tsahkadsor, 1971). Akadjemiai Kiado, Budapest; 1973. p. 267–281.

94. Schwarz G. Estimating the dimension of a model. Ann Statist. 1978;6(2):461–464.

95. Hannan EJ, Quinn BG. The determination of the order of an autoregression. J Roy Statist Soc Ser B. 1979;41(2):190–195.

96. Hannan EJ. The estimation of the order of an ARMA process. Ann Statist. 1980;8(5):1071–1081.

97. Mantegna RN. Lȷevy walks and enhanced diffusion in Milan stock exchange. Physica A. 1991;179:232–242.

Figure captions

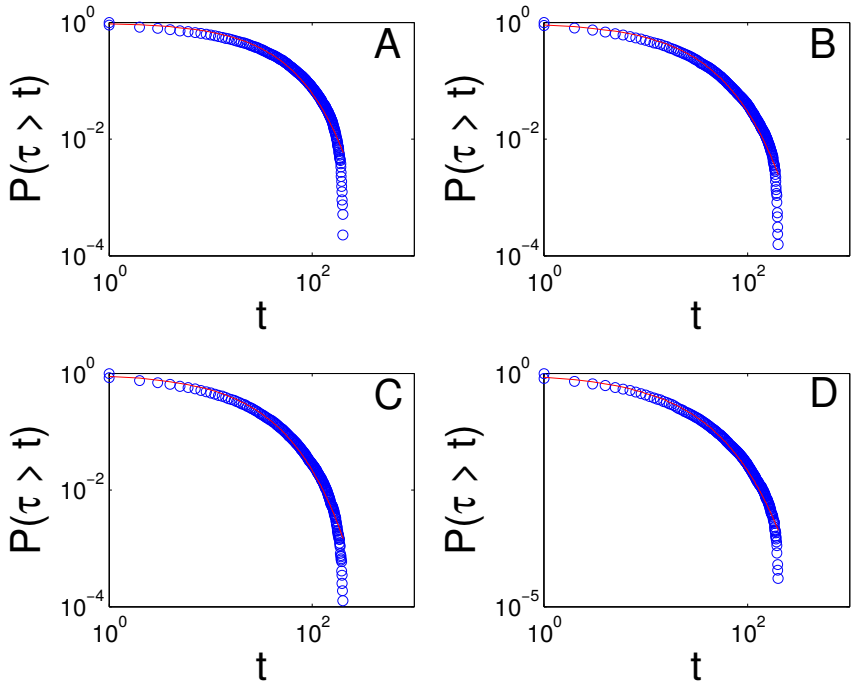

Fig 1. Weibull fit for A2A (A), EXO (B), MS (C), TIT (D). The fit is represented by the thin solid line, the open circles are the empirical values for the survival function.

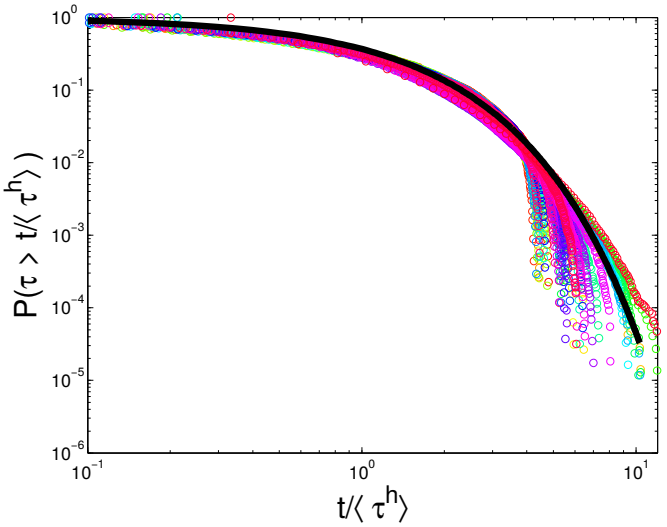

Fig 2. Approximate scaling of the survival function for the forty time series. The solid line is the Weibull fit given by Eq.(3).

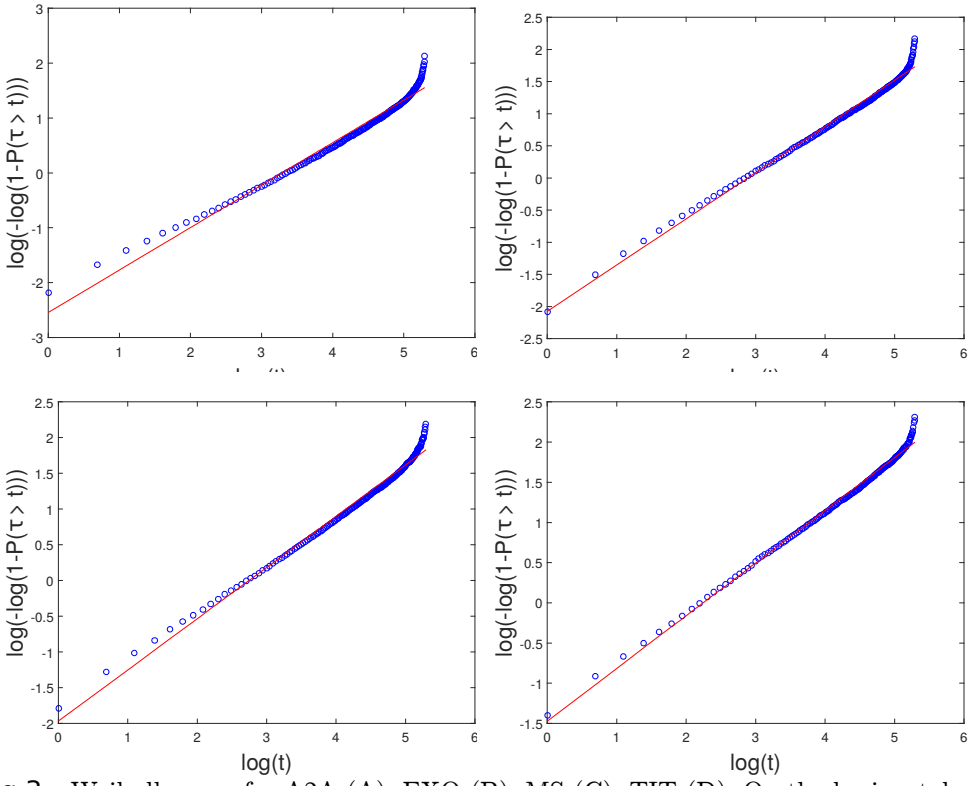

Fig 3. Weibull paper for A2A (A), EXO (B), MS (C), TIT (D). On the horizontal axis, the values of $\log(t)$ are plotted, where t represents the inter-trade duration. On the vertical axes, a double logarithmic transform of the empirical cumulative distribution function of the inter-trade durations is plotted: $\log(-\log(1 - P(\tau > t\ )))$. The linear fit is represented by the thin red solid line, the open circles are the empirical values.

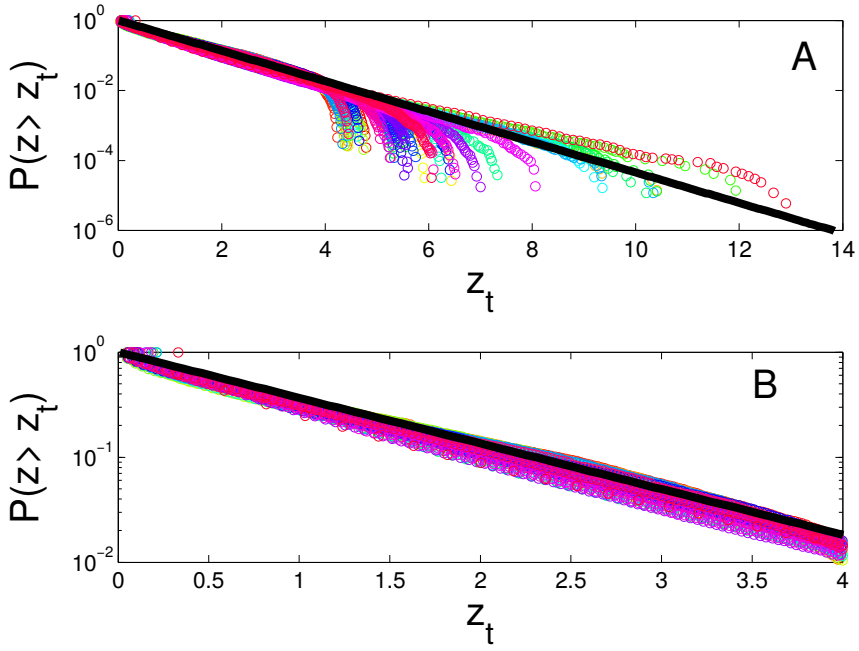

Fig 4. (A) Empirical survival function for the transformed variable z_t^h compared with the expected exponential function $\exp(-z_t^h)$; (B) Zoom in the region $z_t^h \leq 4$.

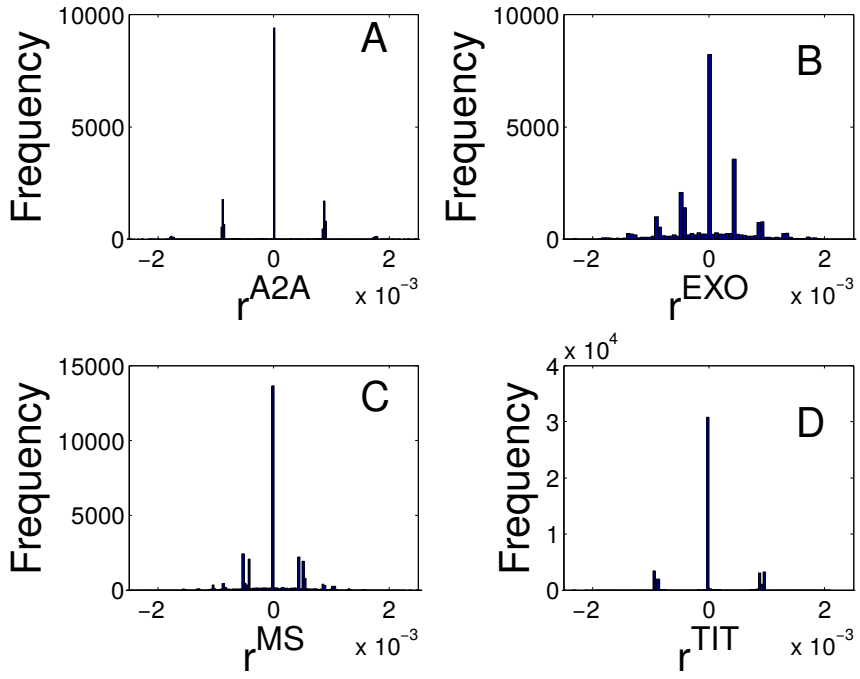

Fig 5. Histogram of returns for A2A (A), EXO (B), MS (C), TIT (D).

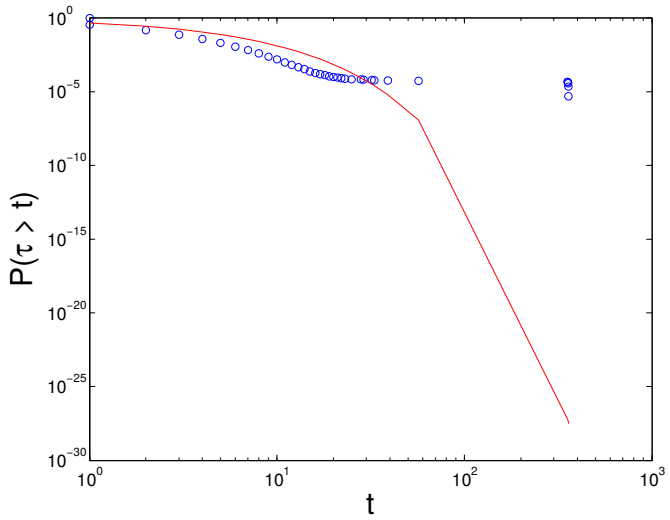

Fig 6. Circles: empirical survival function; solid line: Weibull fit.

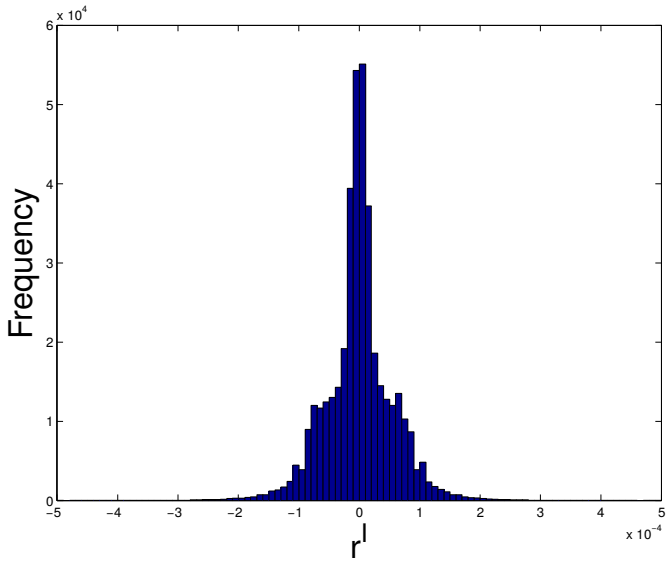

Fig 7. Histogram of returns for the FTSE MIB index.

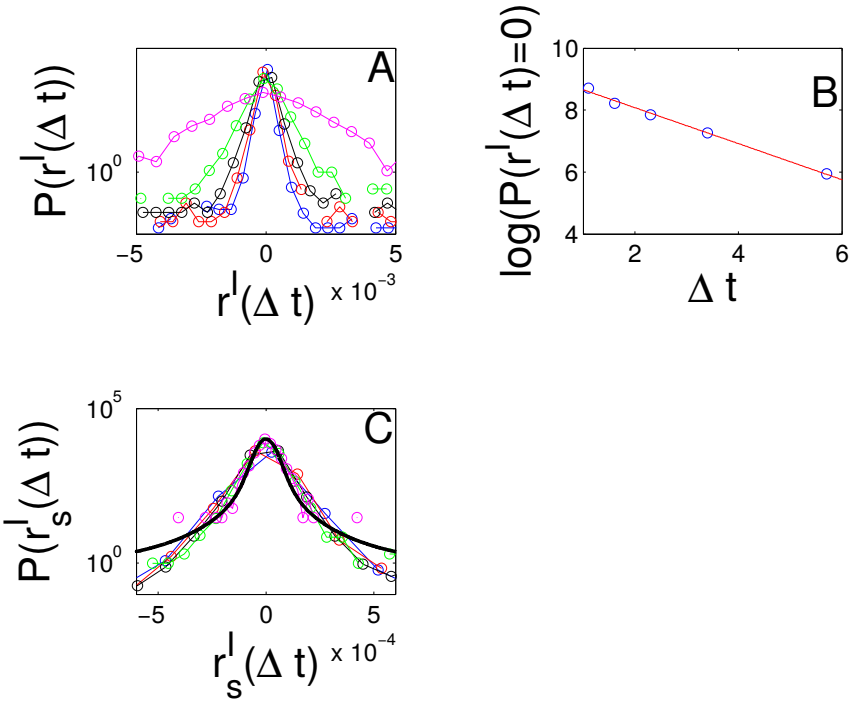

Fig 8. (A) Histogram of the returns for the FTSE MIB index observed at different time intervals, namely, $\Delta t = 3$ s (blue), 5 s (red), 10 s (black), 30 s (green) and 300 s (purple); (B) Probability of zero returns as a function of the time sampling interval Δt, the slope of the straight line is 0.58 ± 0.01; (C) scaled empirical probability distribution and comparison with the theoretical prediction given by Eq.(7) (black solid line).

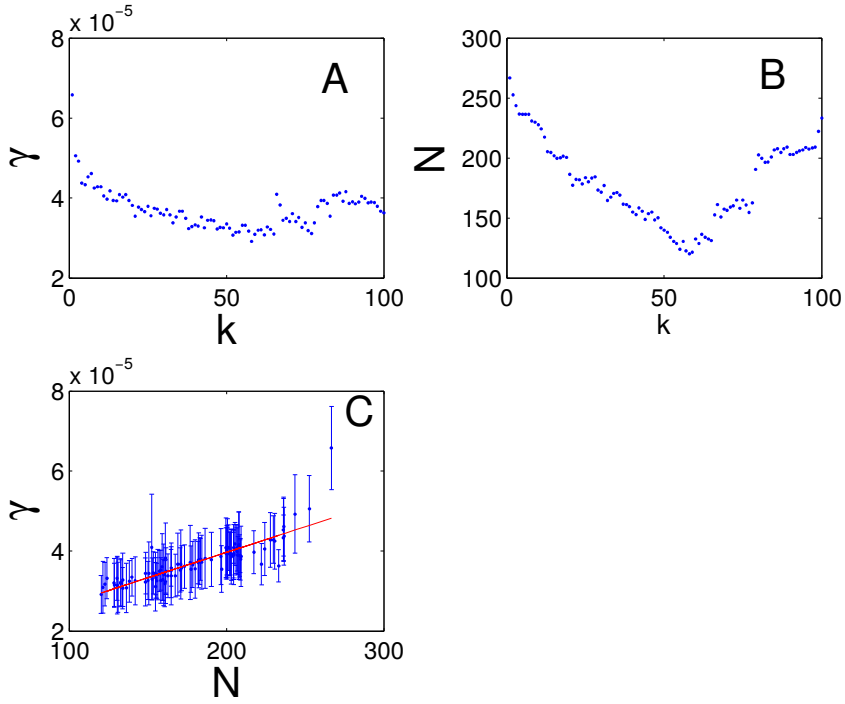

Fig 9. (A) Volatility γ as a function of k for $\delta t = 300$ s. (B) Activity N as a function of k for $\delta t = 300$ s. (C) Scatter plot of volatility γ as a function of number of trades N. The points are averaged over the investigated period.

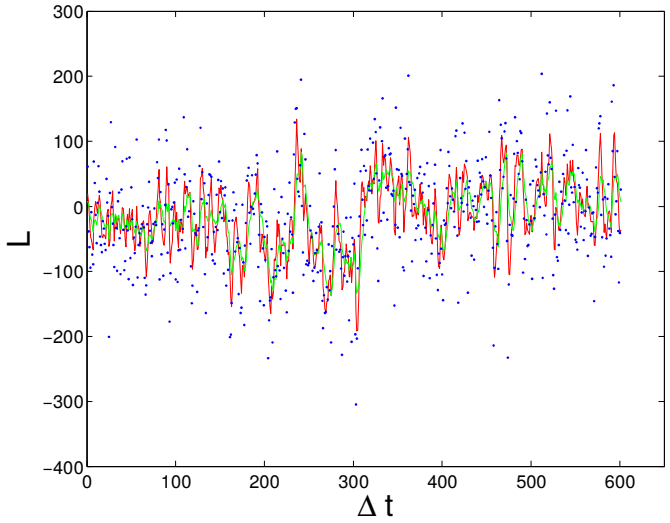

Fig 10. Leverage L as a function of lag Δt. The red and green solid lines show the leading short (4 lags) and lagging long (10 lags) square-root weighted moving average, respectively. Δt is equal to 3s. There is no strong evidence of leverage effect.

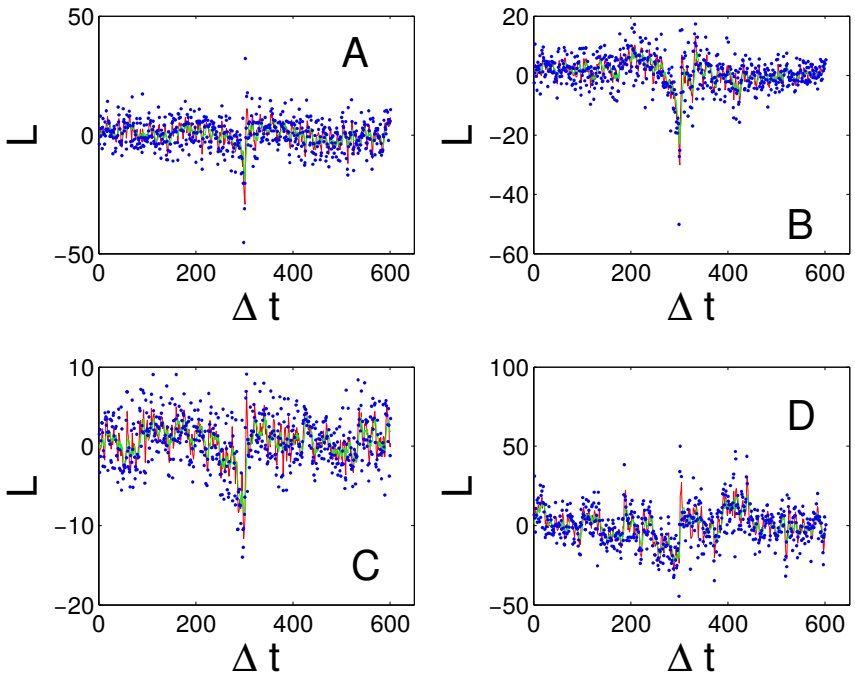

Fig 11. Leverage L as a function of lag Δt for S&P500 index (A), NASDAQ index (B), CAC40 index (C), FTSE index (D). The red and green solid lines show the leading short (4 lags) and lagging long (10 lags) square-root weighted moving average, respectively. The Δt is equal to one day. For S&P500, NASDAQ indices, the leverage effect is well evident, whereas for CAC40 and FTSE indices it is less evident.

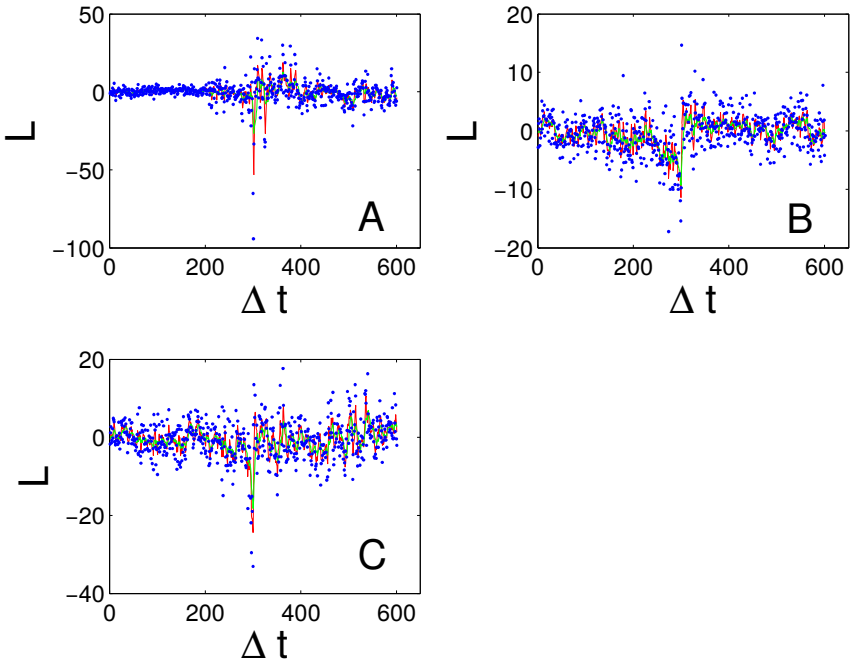

Fig 12. Leverage L as a function of lag Δt for DAX index (A), Nikkei index (B), Hang Seng index (C). The red and green solid lines show the leading short (4 lags) and lagging long (10 lags) square-root weighted moving average, respectively. The Δt is equal to one day. For DAX, Nikkei and Hang Seng indices, the leverage effect is well evident.

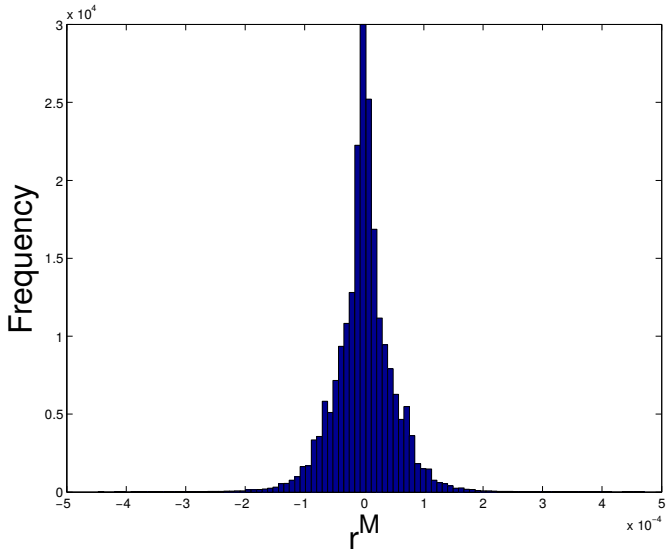

Fig 13. Histogram of returns for the approximating process with $w = 3$s.

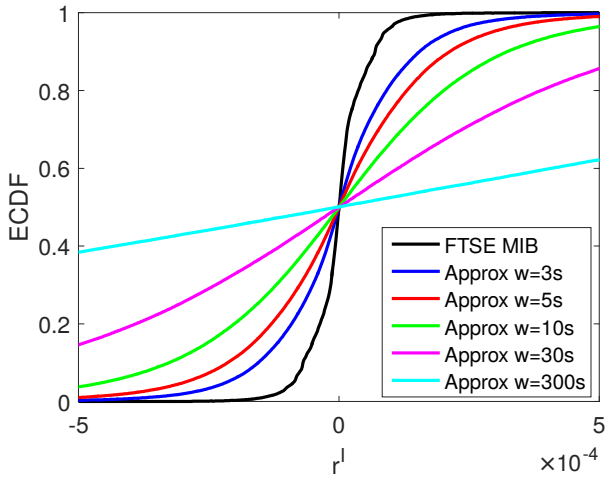

Fig 14. Approximation of the empirical cumulative distribution function for FTSE MIB returns r^{l} .

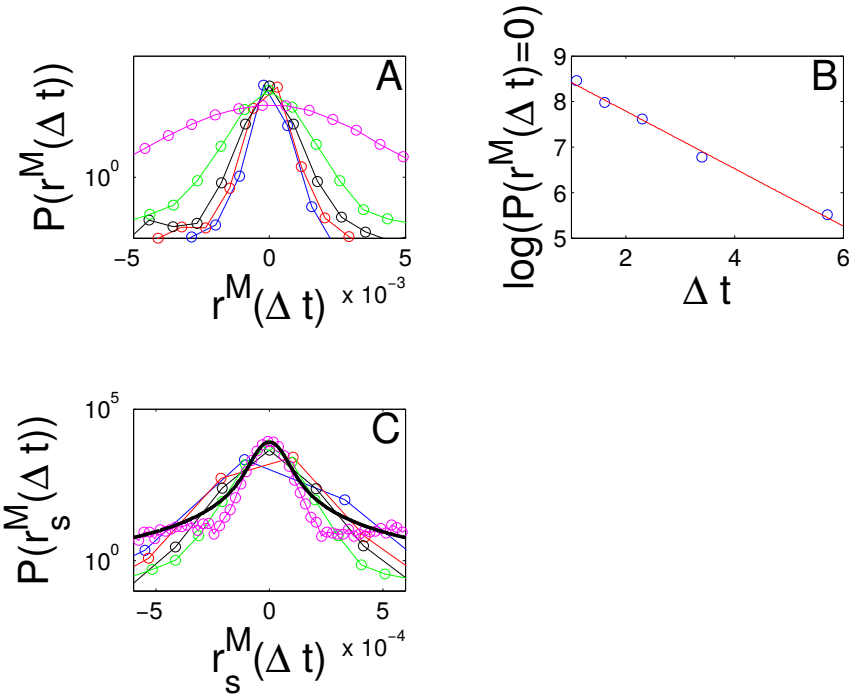

Fig 15. (A) Histogram of the returns for the simulation described in the text observed at different time intervals, namely, $\Delta t = 3$ s (blue), 5 s (red), 10 s (black), 30 s (green) and 300 s (purple); (B) Probability of zero returns as a function of the time sampling interval Δt, the slope of the straight line is 0▷63 ∘ 0▷01; (C) scaled empirical probability distribution and comparison with the theoretical prediction given by Eq.(7) (black solid line).

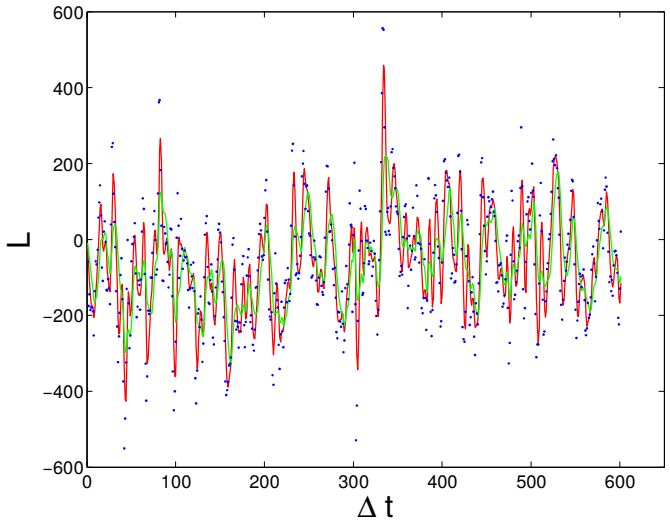

Fig 16. Leverage L as a function of lag Δt for simulated data.The red and green solid lines show the leading short (4 lags) and lagging long (10 lags) square-root weighted moving average, respectively. Δt is equal to 3s. Also for the simulation there is no strong evidence of leverage effect.

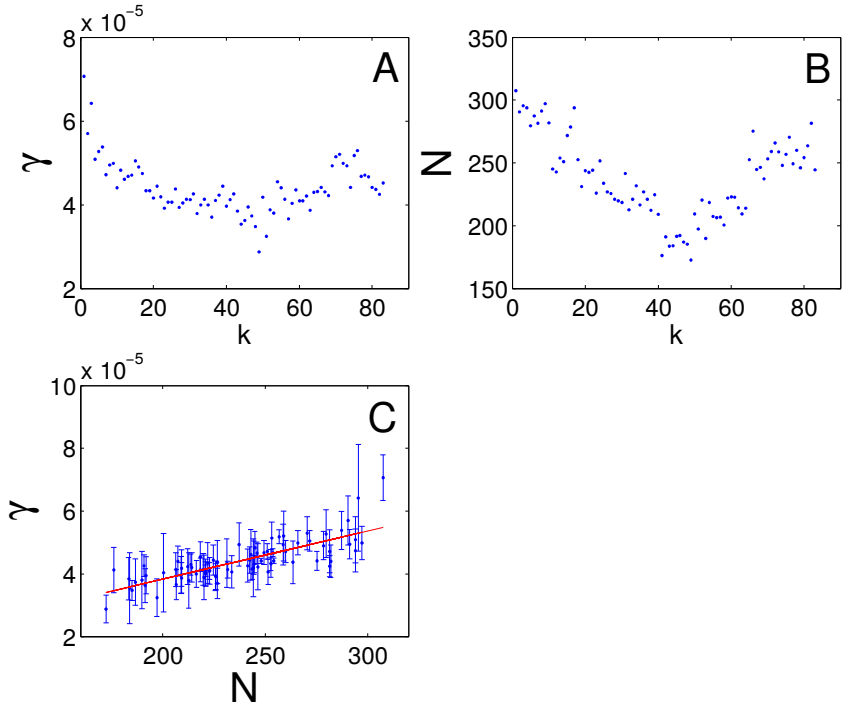

Fig 17. (A) Volatility γ as a function of k for $\delta t = 300$ s. (B) Activity N as a function of k for $\delta t = 300$ s . (C) Scatter plot of volatility γ as a function of number of trades N. The points are averaged over the investigated period. All the plots are for simulated data with $w = 10$ s.

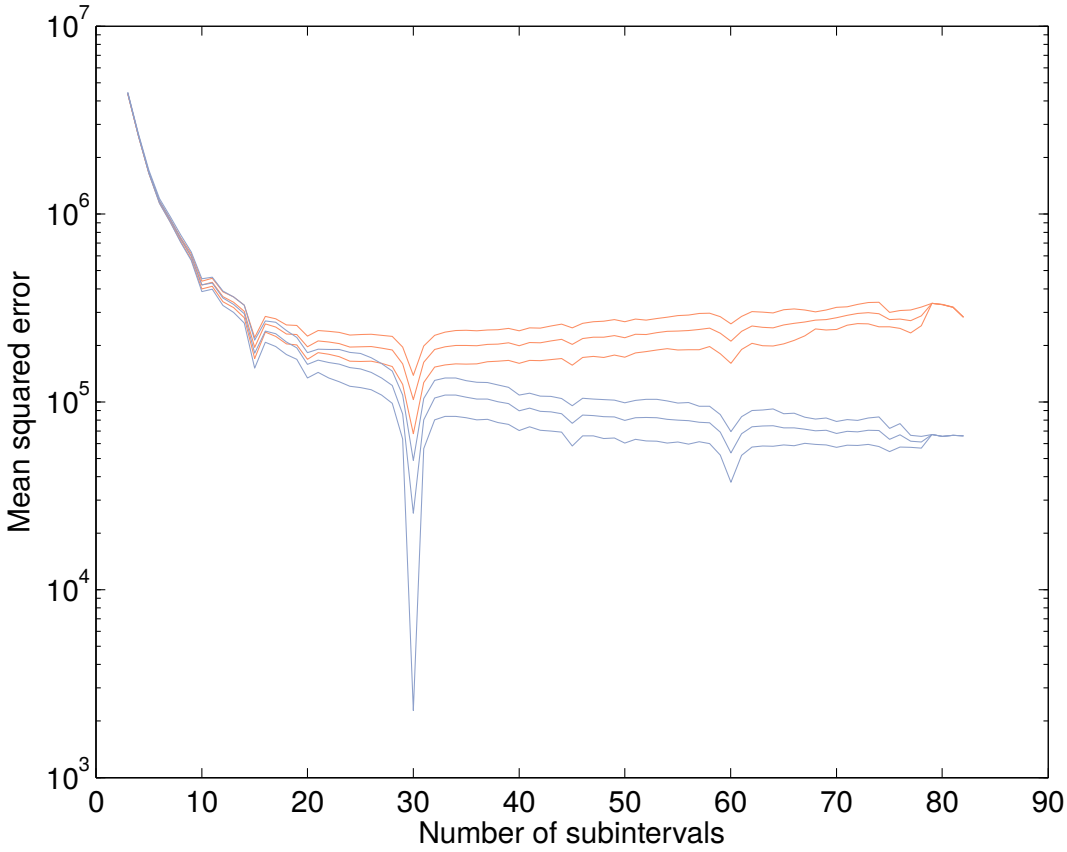

Fig 18. Plot of the mean squared error (MSE) of the estimation of the intensity function for the (Dλ) model (orange lines) and for the (Pλ) model (blue lines) respectively. The graph shows the MSE together with dashed lines indicating the size of the first standard deviation from the mean as a function of the underlying number of intervals of the fitting grid. The true model for the simulation originally used 30 subintervals. The MSE is calculated as a squared L^2 distance between the estimated and the true intensity function (see also Eq. (34)).

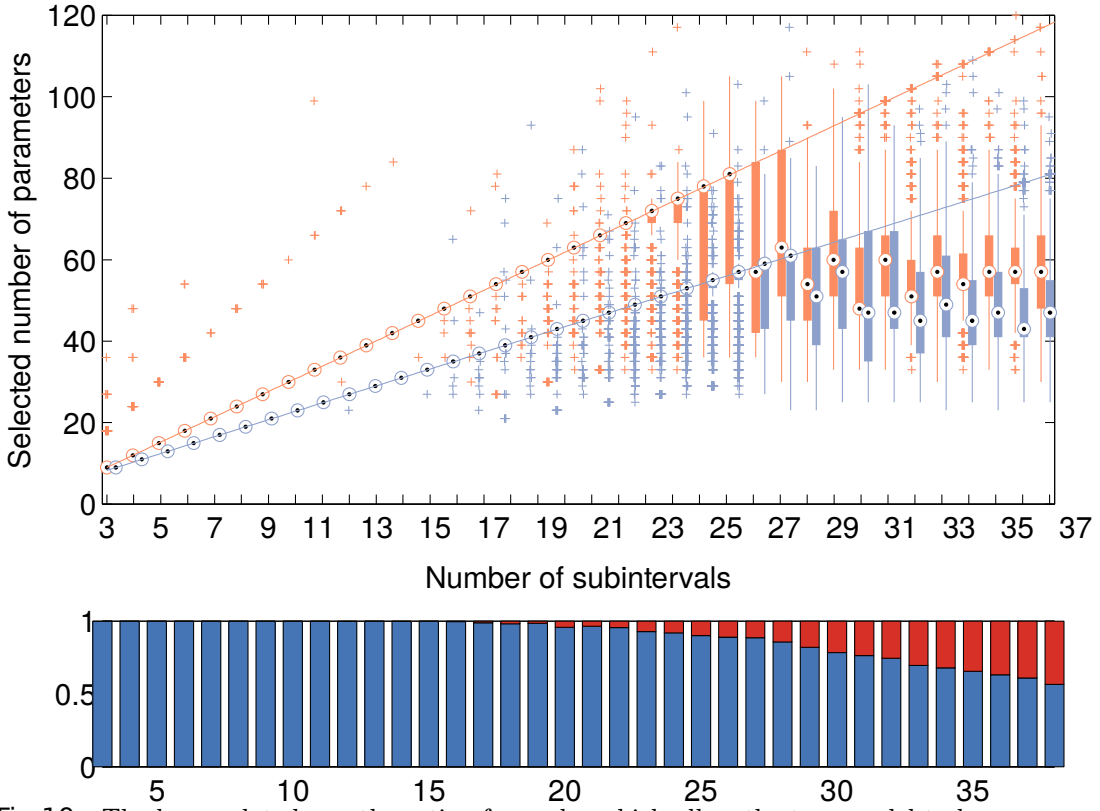

Fig 19. The lower plot shows the ratio of samples which allow the true model to be among the set of models from which the IC may choose from, in other words there is no misspecification (blue areas). This ratio decreases and for finer discretization there are more cases of model selection under misspecification (red areas). The sum of blue and red areas is 100%.
The upper plot shows that the model selection using the AIC for the (Dλ)-model (orange box plot) closely follows the reference line indicating $3n$ (n = number of subintervals) for small n, before deviating for larger n. The same holds for the (Pλ)-model (blue box plot) and its corresponding reference line $2n + 1$. The number of subintervals for which both box plots deviate from their respective reference lines is around $n = 25$ to $n = 27$. In the region $n <$ 15, there are several outliers which are almost all overestimates.

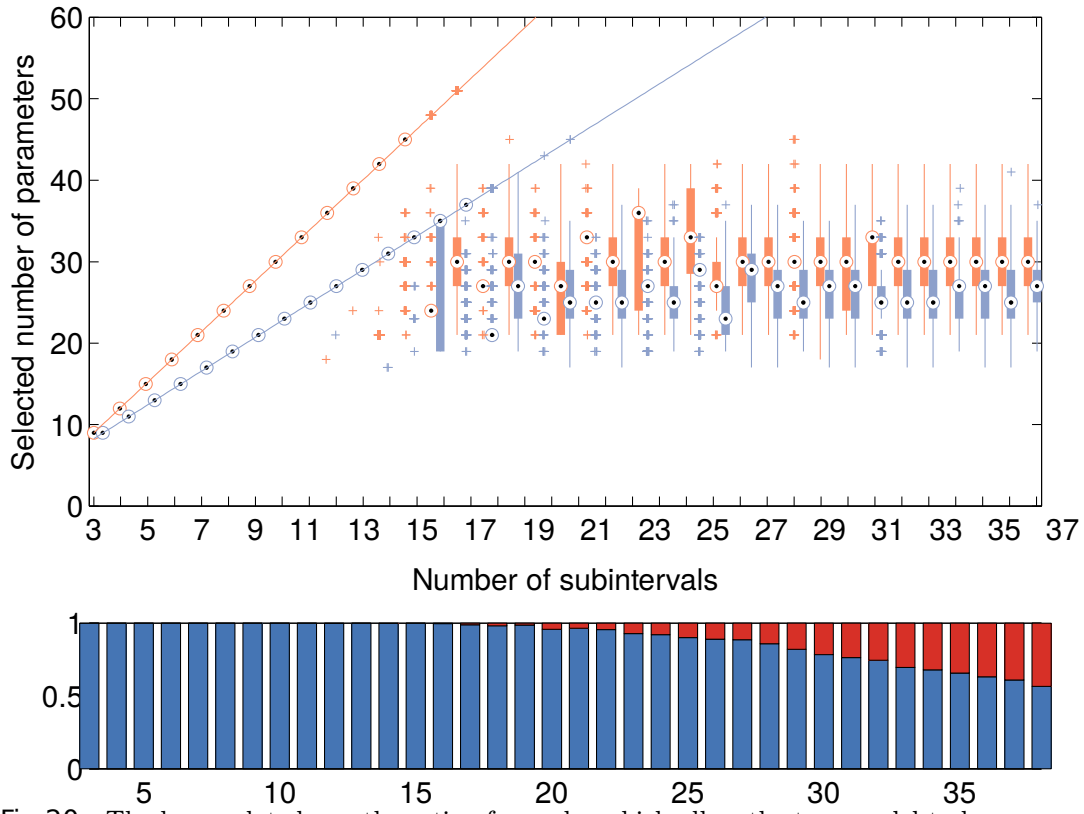

Fig 20. The lower plot shows the ratio of samples which allow the true model to be among the set of models from which the IC may choose from, in other words there is no misspecification (blue areas). This ratio decreases and for finer discretization there are more cases of model selection under misspecification (red areas). The sum of blue and red areas is 100%.
The upper plot shows that the model selection using the BIC for the (Dλ)-model (orange box plot) closely follows the reference line indicating $3n$ (n = number of subintervals) for small n before deviating for larger n. The same holds for the (Pλ)-model (blue box plots) and its corresponding reference line $2n + 1$. The number of subintervals for which both box plots deviate from their respective reference lines is around $n = 15$ to $n = 17$.

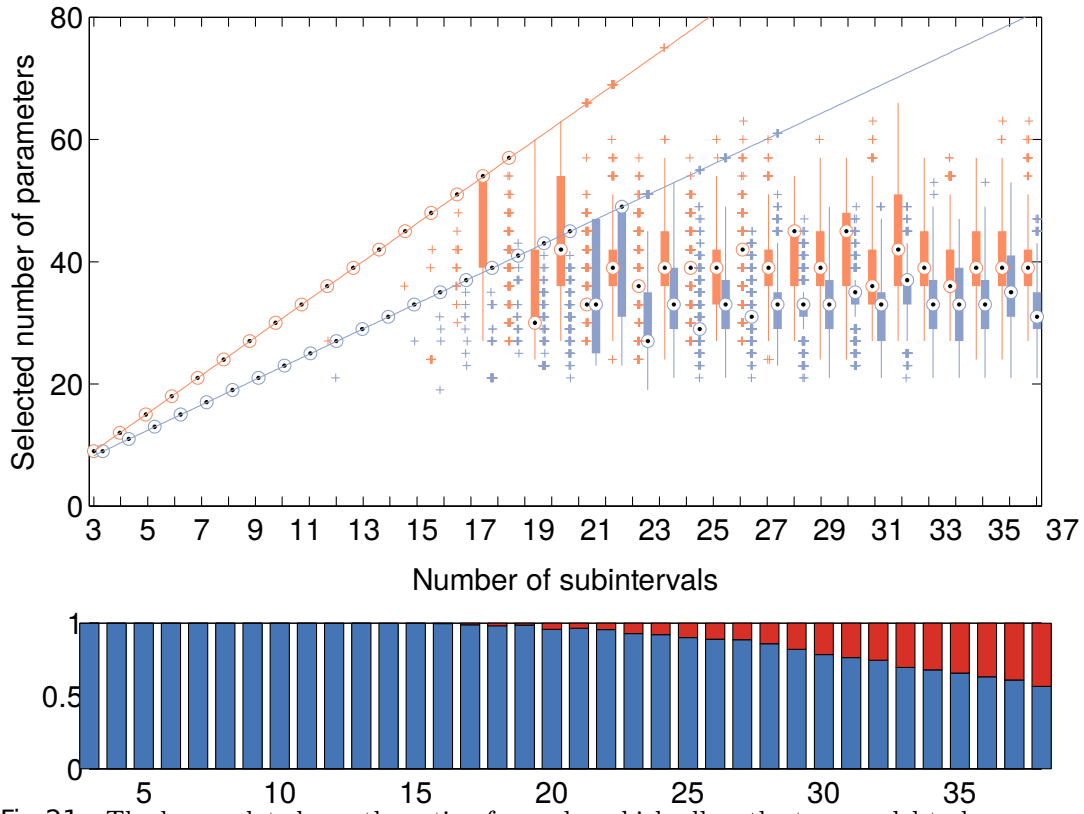

Fig 21. The lower plot shows the ratio of samples which allow the true model to be among the set of models from which the IC may choose from, in other words there is no misspecification (blue areas). This ratio decreases and for finer discretization there are more cases of model selection under misspecification (red areas). The sum of blue and red areas is 100%.
The upper plot shows that the model selection using the HQ for the (Dλ)-model (orange box plot) closely follows the reference line indicating $3n$ (n = number of subintervals) for small n before deviating for larger n. The same holds for the (Pλ)-model (blue box plots) and its corresponding reference line $2n + 1$. The number of subintervals for which both box plots deviate from their respective reference lines is around $n = 18$ to $n = 20$.

Tables

Table 1. Symbols and number of observations for the 40 assets composing the FTSE MIB index in February-March 2011

Asset	Symbol	Number of observations
A2A	A2A	17987
Ansaldo STS	STS	14252
Atlantia	ATL	25811
Autogrill Spa	AGL	15834
Azimut	AZM	14779
Banco Popolare	BP	70373
Bca MPS	BMPS	38005
Bca Pop Milano	PMI	32132
Bulgari	BUL	20164
Buzzi Unicem	BZU	25236
Campari	CPR	14789
Diasorin	DIA	16386
Enel	ENEL	73223
Enel Green Power	EGPW	29305
ENI	ENI	77280
Exor	EXO	26108
Fiat	F	84641
Fiat Industrial	FI	52212
Finmeccanica	FNC	31566
Fondiaria-SAI	FSA	21169
Generali Ass	G	60561
Impregilo	IPG	16414
Intesa Sanpaolo	ISP	84525
Lottomatica	LTO	14313
Luxottica Group	LUX	25717
Mediaset	MS	32019
Mediobanca	MB	37848
Mediolanum	MED	17185
Parmalat	PLT	30861
Pirelli & C	PC	27023
Prysmian	PRY	32806
Saipem	SPM	57592
Snam Rete Gas	SRG	25324
STMicroelectronics	STM	54515
Telecom Italia	TIT	49576
Tenaris	TEN	36410
Terna	TRN	21836
Tod's	TOD	14811
Ubi Banca	UBI	31541
UniCredit	UCG	168433
Index	FTSE MIB	405560

Table 2. Descriptive statistics for the waiting times τ^h

Asset	mean	std	α	β	AD	Lillie
A2A	32.49	39.04	0.053	0.865	106	0.068
STS	34.07	43.68	0.061	0.818	122	0.083
ATL	24.42	32.48	0.088	0.792	263	0.099
AGL	33.20	41.87	0.059	0.830	145	0.082
AZM	34.67	42.35	0.052	0.853	116	0.074
BP	9.54	12.80	0.189	0.786	1158	0.134
BMPS	17.21	23.96	0.130	0.761	401	0.107
PMI	19.95	27.26	0.111	0.773	293	0.099
BUL	24.87	37.02	0.116	0.717	326	0.123
BZU	22.62	33.71	0.125	0.716	435	0.125
CPR	33.77	42.42	0.058	0.833	174	0.092
DIA	30.21	39.91	0.073	0.797	155	0.091
ENEL	9.19	11.60	0.173	0.829	987	0.123
EGPW	21.16	29.31	0.110	0.764	239	0.094
ENI	8.71	12.21	0.221	0.756	1541	0.148
EXO	22.72	31.16	0.101	0.771	228	0.094
F	7.94	11.29	0.243	0.747	1936	0.158
FI	12.80	18.77	0.182	0.726	833	0.132
FNC	20.86	26.98	0.093	0.812	234	0.089
FSA	23.70	35.15	0.120	0.719	309	0.118
G	11.10	14.79	0.165	0.792	759	0.119
IPG	32.26	41.41	0.064	0.818	157	0.085
ISP	7.96	11.30	0.242	0.748	1930	0.158
LTO	33.22	42.54	0.062	0.819	117	0.082
LUX	23.28	31.52	0.096	0.780	231	0.096
MS	20.12	27.93	0.114	0.763	350	0.107
MB	17.40	24.03	0.126	0.767	403	0.108
MED	31.66	39.57	0.060	0.837	126	0.077
PLT	20.49	29.01	0.119	0.749	322	0.104
PC	22.78	30.45	0.094	0.789	221	0.092
PRY	19.48	27.87	0.126	0.743	390	0.113
SPM	11.53	17.88	0.219	0.691	1185	0.150
SRG	24.77	32.77	0.086	0.796	208	0.091
STM	12.22	17.26	0.174	0.751	750	0.124
TIT	13.27	20.52	0.198	0.692	972	0.146
TEN	17.49	24.98	0.137	0.743	395	0.110
TRN	28.12	35.52	0.068	0.829	148	0.080
TOD	31.31	40.71	0.068	0.808	114	0.081
UBI	20.58	27.30	0.100	0.794	272	0.096
UCG	3.85	4.94	0.364	0.817	8640	0.223
Index	1.66	1.26	–	–	Inf	0.365

Table 3. Descriptive statistics for the volumes v^{h}

Assets	mean $\times 10^{0}$	variance $\times 10^{0}$	skewness	kurtosis $\times 10^{0}$
A2A	1.11	5.72	11.17	2.75
STS	0.11	0.05	10.86	2.79
ATL	0.16	0.09	8.79	2.16
AGL	0.15	0.09	7.97	1.26
AZM	0.13	0.05	6.10	0.70
BP	1.17	6.21	20.98	12.14
BMPS	1.69	10.05	6.98	1.01
PMI	0.52	0.67	5.64	0.74
BUL	0.53	7.33	26.99	13.21
BZU	0.16	0.07	7.05	1.21
CPR	0.18	0.08	5.66	0.61
DIA	0.03	0.28×10^{2}	6.33	0.73
ENEL	1.09	7.06	15.92	6.97
EGPW	0.80	2.78	12.88	3.60
ENI	0.48	2.20	78.73	118.40
EXO	0.07	0.01	5.10	0.49
F	0.62	1.68	9.31	2.05
FI	0.31	0.37	8.04	1.36
FNC	0.18	0.14	10.76	3.01
FSA	0.22	0.14	10.39	3.42
G	0.31	0.35	9.09	2.32
IPG	0.56	1.40	13.44	3.88
ISP	3.39	45.25	7.56	1.72
LTO	0.14	0.07	6.67	0.81
LUX	0.08	0.02	10.30	2.83
MS	0.41	0.54	7.23	1.19
MB	0.28	0.26	8.41	1.66
MED	0.31	0.33	10.09	2.29
PLT	1.01	8.72	31.52	17.87
PC	0.33	0.37	9.07	2.07
PRY	0.14	0.07	7.80	1.32
SPM	0.09	0.03	13.07	5.58
SRG	0.56	6.92	117.04	166.34
STM	0.29	0.32	7.21	1.29
TIT	3.26	66.70	18.81	11.30
TEN	0.17	0.09	9.18	2.11
TRN	0.83	5.89	61.35	64.82
TOD	0.02	0.08×10^{2}	7.52	1.07
UBI	0.23	0.17	6.87	1.01
UCG	5.63	124.95	7.82	1.78

Table 4. Descriptive statistics for the trade-by-trade log-returns r^h. (*) On March 7th, 2011, the French firm LVMH launched a takeover offer (OPA - Offerta Pubblica d'Acquisto in Italian) to buy Bulgari shares at 12.25 euros. On that day, this share price jumped from below 8 euros to more than 12 euros.

Assets	mean $\times 10^{-7}$	variance $\times 10^{-7}$	skewness $\times 10^{-2}$	kurtosis
A2A	29.15	5.24	9.36	5.22
STS	−14.43	6.76	−7.11	11.50
ATL	1.59	2.09	24.62	19.64
AGL	−36.50	6.09	114.90	43.47
AZM	−3.29	8.03	−21.90	14.14
BP	−4.53	4.55	−1.69	10.69
BMPS	24.93	4.79	−21.71	24.34
PMI	6.87	5.55	−23.73	41.72
BUL (*)	−3.75	4.37	−295.68	154.69
BZU	61.92	7.41	−99.04	35.92
CPR	2.35	3.73	11.04	8.13
DIA	−40.04	4.42	−49.99	29.17
ENEL	6.21	1.38	140.10	76.06
EGPW	38.81	3.64	3.43	7.31
ENI	7.86	1.40	59.89	21.01
EXO	11.98	4.82	−5.45	8.06
F	−3.55	2.81	−45.05	21.76
FI	14.33	3.68	−39.37	18.14
FNC	0.50	3.29	28.01	13.01
FSA	84.68	10.35	−163.51	180.64
G	5.03	2.09	−100.65	44.97
IPG	80.66	9.04	−45.81	22.68
ISP	1.99	3.45	−62.87	43.12
LTO	67.82	9.28	−171.44	62.62
LUX	25.88	2.67	30.48	24.43
MS	5.76	2.86	−22.98	19.38
MB	17.29	4.18	1.66	9.67
MED	20.25	7.64	−43.78	18.78
PLT	9.76	5.30	49.56	14.43
PC	47.93	5.41	3.44	10.75
PRY	21.54	4.02	257.09	92.76
SPM	5.72	1.50	−9.12	32.75
SRG	12.09	2.41	79.03	54.87
STM	15.69	2.56	−39.64	36.78
TIT	8.33	3.20	−22.22	8.92
TEN	0.34	2.61	−112.99	135.05
TRN	26.67	2.42	3.54	6.03
TOD	28.73	6.95	158.96	86.49
UBI	−1.76	4.99	−67.53	25.23
UCG	3.44	1.29	−12.56	57.51
Index	1.10	0.03	2	8.54

Table 5. Kolmogorov-Smirnov test, Jarque-Bera test and Lilliefors test to test the nomality of the empirical distributions corresponding to Δt equal to 3s, 5s, 10s, 30s, 300s. The null hypothesis is always rejected.

Δt	K-S test	J-B test	Lilliefors test
3s	0.499	3108277492760.052	0.135
5s	0.499	1087100884817.007	0.142
10s	0.499	117920812948.739	0.141
30s	0.498	3409200688.215	0.128
300s	0.497	7949646.601	0.136

Table 6. Kolmogorov-Smirnov test. In bold the rejected cases. The null hypothesis of empirical data coming from an identical distribution is rejected in the comparisons of $\Delta t = 3s$ and $\Delta t = 5s$, $\Delta t = 3s$ and $\Delta t = 10s$ and $\Delta t = 3s$ and $\Delta t = 30s$.

Δt	3s	5s	10s	30s	300s
3s	–	0.010	0.014	0.014	0.023
5s	0.010	–	0.008	0.010	0.022
10s	0.014	0.008	–	0.008	0.017
30s	0.014	0.010	0.008	–	0.018
300s	0.023	0.022	0.017	0.018	–

Table 7. Kolmogorov-Smirnov test, Jarque-Bera test and Lilliefors test for the normality of simulated data

Δt	K-S test	J-B test	Lilliefors test
3s	0.499	12727206295498.855	0.209
5s	0.499	2484970052007.152	0.200
10s	0.499	279530930024.888	0.189
30s	0.498	9268127864.106	0.173
300s	0.490	8190272.873	0.157

Table 8. Kolmogorov-Smirnov test. The null hypothesis of simulated data coming from an identical distribution is always rejected.

Δt	3s	5s	10s	30s	300s
3s	–	0.019	0.031	0.036	0.035
5s	0.019	–	0.012	0.018	0.018
10s	0.031	0.012	–	0.007	0.016
30s	0.036	0.018	0.007	–	0.019
300s	0.035	0.018	0.016	0.019	–

Table 9. Parameter settings for the simulation of ACD data

	ω	α_1	α_2	β_1	β_2
ACD(1,1)	1	0.089	–	0.85	–
ACD(1,2)	1	0.1	–	0.45	0.4
ACD(2,1)	1	0.15	0.15	0.65	–
ACD(2,2)	1	0.1	0.1	0.42	0.35

Table 10. Table of summary statistics of the MSE of the parameters θ and σ^2 of the compound Poisson type model. The analysis is based on 1000 samples generated from a simulation grid containing 30 subintervals.

	mean	min	max	std
θ	0.0545	0.0026	0.1049	0.0212
σ^2	0.1038	0.0049	0.1757	0.0439

Table 11. Results of the MSE calculations for the ACD model

		MSE(ω)	MSE(α_1)	MSE(α_2)	MSE(β_1)	MSE(β_2)
ACD(1,1)	T=250	3.7508	0.0023	–	0.0231	–
	T=500	1.8887	0.0010	–	0.0108	–
	T=1000	0.3591	0.0005	–	0.0025	–
	T=2000	0.1245	0.0002	–	0.0010	–
ACD(1,2)	T=250	14.5255	0.0036	–	0.4748	0.4282
	T=500	3.7468	0.0019	–	0.3039	0.2681
	T=1000	0.6259	0.0010	–	0.1869	0.1606
	T=2000	0.1905	0.0005	–	0.0809	0.0681
ACD(2,1)	T=250	0.8491	0.0063	0.0108	0.0130	–
	T=500	0.2664	0.0032	0.0050	0.0053	–
	T=1000	0.0916	0.0014	0.0026	0.0023	–
	T=2000	0.0418	0.0007	0.0012	0.0011	–
ACD(2,2)	T=250	6.4135	0.0067	0.0102	0.3165	0.2445
	T=500	1.1077	0.0032	0.0061	0.2722	0.2031
	T=1000	0.3730	0.0014	0.0041	0.2086	0.1526
	T=2000	0.1512	0.0006	0.0026	0.1612	0.1181

Table 12. Model selection results based on ACD(1,1) data samples: Given 1000 samples of size $T \in$¶250‹ 500‹ 1000‹ 2000◊each column gives the percentage of cases in which the different IC selected the models ACD(1,1), ACD(1,2), ACD(2,1) and ACD(2,2) respectively. The bold numbers give the largest percentage per row.

		ACD(1,1)	ACD(1,2)	ACD(2,1)	ACD(2,2)
T=250	AIC	**58.7**	23.6	9.9	7.8
	BIC	**90.2**	7	2.1	0.7
	HQ	**77.9**	14.6	4.8	2.7
T=500	AIC	**62.9**	20.4	10.9	5.8
	BIC	**93.6**	4.7	1.6	0.1
	HQ	**82.6**	11.5	4.9	1
T=1000	AIC	**67.5**	16.4	11	5.1
	BIC	**97.4**	1.8	0.8	0
	HQ	**87.2**	7.5	4.8	0.5
T=2000	AIC	**71.3**	13.1	9.7	5.9
	BIC	**97.7**	1.6	0.6	0.1
	HQ	**91.5**	4.4	3	1.1

Table 13. Model selection results based on ACD(1,2) data samples: Given 1000 samples of size $T \in$¶250‹ 500‹ 1000‹ 2000◊each column gives the percentage of cases in which the different IC selected the models ACD(1,1), ACD(1,2), ACD(2,1) and ACD(2,2) respectively. The bold numbers give the largest percentage per row.

		ACD(1,1)	ACD(1,2)	ACD(2,1)	ACD(2,2)
T=250	AIC	**58.6**	24.7	9.6	7.1
	BIC	**91.5**	6.5	1.3	0.7
	HQ	**78.6**	14.8	3.7	2.9
T=500	AIC	**60.6**	25.1	10.3	4
	BIC	**94.7**	4.3	0.7	0.3
	HQ	**81.2**	13.5	4.5	0.8
T=1000	AIC	**52.7**	27.8	15.2	4.3
	BIC	**92.6**	5.1	2.3	0
	HQ	**76**	14.7	8.8	0.5
T=2000	AIC	**41.5**	35.6	18	4.9
	BIC	**88.4**	6.7	4.9	0
	HQ	**67.6**	20.4	11.6	0.4

Table 14. Model selection results based on ACD(2,1) data samples: Given 1000 samples of size $T \in$¶250‹ 500‹ 1000‹ 2000◊each column gives the percentage of cases in which the different IC selected the models ACD(1,1), ACD(1,2), ACD(2,1) and ACD(2,2) respectively. The bold numbers give the largest percentage per row.

		ACD(1,1)	ACD(1,2)	ACD(2,1)	ACD(2,2)
T=250	AIC	**36.2**	20.9	31.8	11.1
	BIC	**73.7**	8.9	16.8	0.6
	HQ	**52.4**	16.3	28.1	3.2
T=500	AIC	19.1	20.7	**50**	10.2
	BIC	**59.9**	10.5	29	0.6
	HQ	36.5	16.4	**43.8**	3.3
T=1000	AIC	7.4	16.7	**64.8**	11.1
	BIC	35.6	11.9	**52.1**	0.4
	HQ	17.1	15.7	**63.7**	3.5
T=2000	AIC	1.2	12.7	**74.2**	11.9
	BIC	6.8	12.9	**80.1**	0.2
	HQ	2.2	14.2	**81.6**	2

Table 15. Model selection results based on ACD(2,2) data samples: Given 1000 samples of size $T \in$¶250‹ 500‹ 1000‹ 2000◊each column gives the percentage of cases in which the different IC selected the models ACD(1,1), ACD(1,2), ACD(2,1) and ACD(2,2) respectively. The bold numbers give the largest percentage per row.

		ACD(1,1)	ACD(1,2)	ACD(2,1)	ACD(2,2)
T=250	AIC	**56.7**	15.8	18.8	8.7
	BIC	**89.7**	5.3	4.5	0.5
	HQ	**74**	11.5	11.7	2.8
T=500	AIC	**57.2**	13.6	19.1	10.1
	BIC	**92.1**	2.9	4.6	0.4
	HQ	**78.4**	8	11.4	2.2
T=1000	AIC	**48.4**	13.1	23.4	15.1
	BIC	**91.5**	2.7	5.7	0.1
	HQ	**74**	6.9	16.1	3
T=2000	AIC	34.2	9.7	**37.2**	18.9
	BIC	**86.1**	1.8	11.5	0.6
	HQ	**59.7**	6.8	26.5	7